Irreducibility of nonsmooth state-space models with an application to CMA-ES

Armand Gissler, Shan-Conrad Wolf, Anne Auger, Nikolaus Hansen

Irreducibility of nonsmooth state-space models with an application to CMA-ES

Armand Gissler, Shan-Conrad Wolf, Anne Auger, Nikolaus Hansen

September 30, 2024

Abstract

We analyze a stochastic process resulting from the normalization of states in the zeroth-order optimization method CMA-ES. On a specific class of minimization problems where the objective function is scaling-invariant, this process defines a time-homogeneous Markov chain whose convergence at a geometric rate can imply the linear convergence of CMA-ES. However, the analysis of the intricate updates for this process constitute a great mathematical challenge. We establish that this Markov chain is an irreducible and aperiodic T-chain. These contributions represent a first major step for the convergence analysis towards a stationary distribution. We rely for this analysis on conditions for the irreducibility of nonsmooth state-space models on manifolds. To obtain our results, we extend these conditions to address the irreducibility in different hyperparameter settings that define different Markov chains, and to include nonsmooth state spaces.

Contents

1 Introduction

The convergence of stochastic processes is at the core of many algorithms in various domains. Well-known examples include Markov chain Monte-Carlo (MCMC) algorithms [11] like the Metropolis-Hastings algorithm [33, 28] that aim to sample a target distribution π by generating a Markov chain with stationary probability measure π. Fast convergence of the Markov chain towards π is one important property for the underlying algorithms. It can be described qualitatively as the geometric ergodicity of the Markov chain, i.e., convergence at a geometric rate towards π, a question that has been widely studied [17, 38]. We focus here on an application of stochastic processes to the domain of numerical stochastic optimization which is closely connected to MCMC. We analyze indeed a Markov chain underlying the so-called covariance matrix adaptation evolution strategy (CMA-ES) [26, 23], a widely used stochastic derivative-free optimization algorithm [39, 14, 9, 16, 2, 32, 41, 1][1] that can tackle difficult optimization problems which are notably nonconvex, multimodal and ill-conditioned. The algorithm minimizes a function $f : \mathbb{R}^n \to \mathbb{R}$ by sampling Gaussian vectors whose mean and covariance matrix are adapted iteratively. The adaptation of the parameters of the Gaussian distribution has been carefully designed, combining several independent principles [25, 26, 24, 37]. Ample empirical evidence shows that the algorithm converges geometrically fast [26, 24, 22, 20]—in optimization referred to as linear convergence—towards the optimum on large classes of functions and the covariance matrix learns the inverse Hessian [21] up to a scalar factor on strictly convex quadratic problems. Yet, establishing a convergence proof of CMA-ES that reflects its working principle (i.e., without modifying the algorithm to enforce convergence) is still an open and difficult theoretical question.

In this context, we extend a methodology that was already successful to analyze stepsize adaptive algorithms [5, 7, 10, 8, 42] to prove the convergence of CMA-ES by exploiting its mechanisms and reflecting its working principle, including the learning of second-order information. The methodology is based on the definition of a normalized Markov chain that models the algorithm when minimizing a scaling-invariant function, a function class that includes non quasi-convex functions [43]. As we will explain, if this Markov chain is stable—in the sense that it converges to a stationary distribution geometrically fast and satisfies a Law of Large Numbers—then the linear convergence of the algorithm follows. With more work, the learning of the inverse Hessian on strictly convex-quadratic functions should follow as well. In order to obtain such stability properties, the irreducibility of the process (the definitions will be recalled in the paper) is a necessary condition. On top of establishing the irreducibility of this Markov chain, we prove that it is an aperiodic T-chain, paving the way to a convergence analysis by means of a geometric drift condition.

Because of the intricacy of the algorithm, the irreducibility cannot be easily established by simply investigating the transition kernel of the Markov chain. Instead, we rely on recent results connecting the irreducibility of a Markov chain defined on a smooth manifold to the stability of an underlying control model. More precisely, we view the Markov chain as a nonlinear state-space model

$$\phi_{t+1} = F(\phi_t, \alpha(\phi_t, U_{t+1})) \tag{1.1}$$

where $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an independent and identically distributed (i.i.d.) process valued in a measured space U, $F\colon \mathsf{X} \times \mathsf{V} \to \mathsf{X}$ is a locally Lipschitz update function between smooth manifolds X, V and $\alpha\colon \mathsf{X} \times \mathsf{U} \to \mathsf{V}$ is a measurable, possibly discontinuous function. When F is nonsmooth, we call (1.1) a nonsmooth state-space model. The connections that we rely on between the irreducibility, aperiodicity and T-chain property of the Markov chain and an underlying deterministic control model have been recently established [19], relaxing the assumptions in previous work [12] that the state space of the chain is an open subset of an Euclidean space and F is continuously differentiable. This latter work was already a generalization of the case where $\alpha(\phi_t, U_{t+1}) = U_{t+1}$ and F is smooth, i.e., infinitely differentiable [34].

While part of the methodology we follow relies on the results presented in [19], we introduce here two other generic and central techniques for the analysis.

Like in many practically used algorithms (in contrast to toy algorithms), different update mechanisms can be turned on and off in CMA-ES by some specific hyperparameter settings (like learning rates) resulting in different algorithm variants with varying number of state variables. Our aim is to analyze all algorithm variants without repeating the similar mathematical analysis for each of them. Hence, in order to have a single proof, we introduce the notions of *projected* and *redundant* Markov chains. Specifically, we consider a Markov chain $\{(\phi_t, \xi_t)\}_{t\in\mathbb{N}}$ valued in the manifold $\mathsf{X} \times \mathsf{Y}$ with

$$(\phi_{t+1}, \xi_{t+1}) = \tilde{F}((\phi_t, \xi_t), \tilde{\alpha}((\phi_t, \xi_t), U_{t+1})) \tag{1.2}$$

where $\{\phi_t\}_{t\in\mathbb{N}}$ obeys (1.1), and $\tilde{F}\colon \mathsf{X} \times \mathsf{Y} \times \mathsf{V} \to \mathsf{X} \times \mathsf{Y}$ and $\tilde{\alpha}\colon \mathsf{X} \times \mathsf{Y} \times \mathsf{U} \to \mathsf{X} \times \mathsf{Y}$ satisfy the same assumptions as

[1] As of March 2024, two Python implementations of CMA-ES received together more than 60 millions downloads.

F and α, respectively. We suppose then that

$$\Pi_{\mathsf{X}} \circ \tilde{F}\left((\phi, \xi), \tilde{\alpha}((\phi, \xi), u)\right) = F(\phi, \alpha(\phi, u)) \tag{1.3}$$

for every $(\phi, \xi, u) \in \mathsf{X} \times \mathsf{Y} \times \mathsf{U}$, where $\Pi_{\mathsf{X}} \colon \mathsf{X} \times \mathsf{Y} \to \mathsf{X}$ is the canonical projection of $\mathsf{X} \times \mathsf{Y}$ on X. The Markov chain $\{(\phi_t, \xi_t)\}_{t \in \mathbb{N}}$ is said to be redundant, whereas $\{\phi_t\}_{t \in \mathbb{N}}$ is said to be projected. We derive similar tools as in [19] to analyze the projected Markov chain $\{\phi_t\}_{t \in \mathbb{N}}$ by investigating the redundant control model (1.2).

Contributions Overall the contributions of this paper are twofold.

On the one hand, we provide two generic tools to analyze the irreducibility, aperiodicity and topological properties of complex nonsmooth state-space models. First, we extend the methodology to investigate Markov chains following (1.1) with locally Lipschitz updates on smooth manifolds in order to be able to deduce irreducibility, aperiodicity and T-chain property from a redundant chain to a projected chain. Second, we show how to transfer the analysis of nonsmooth state-space models following (1.1) from smooth manifolds to *nonsmooth* manifolds, as long as they can be continuously transformed into smooth manifolds.

On the other hand, using the developed tools, we establish the irreducibility, aperiodicity, and T-chain property of a Markov chain defined by the normalization of states of CMA-ES when minimizing a scaling-invariant function. Our results include most of the relevant hyperparameter settings, some of them described by separate Markov chains. The proven properties constitute an essential step for a proof of linear convergence of CMA-ES.

Organization In Section 2, we present the update equations behind CMA-ES and define a class of normalized Markov chains associated to the algorithm when minimizing scaling-invariant functions. In Section 3, we state our first main result that these Markov chains are irreducible, aperiodic T-chains. In Section 4, we state and prove results on the irreducibility, aperiodicity and topological properties of nonlinear state-space models. In Section 5, we apply the results exposed in Section 4 to the normalized Markov chain defined earlier ant prove the main result of Section 3. For the sake of readability, some proofs are delayed and presented in Appendix A and Appendix B.

Notations Throughout this paper, we use the following notations: $\mathbb{N}$, $\mathbb{N}^*$, $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_{++}$ for the sets of nonnegative integers, positive integers, real numbers, nonnegative real numbers, and positive real numbers, respectively. Unless stated otherwise, for $n \in \mathbb{N}^*$ and any vector $x \in \mathbb{R}^n$, $\|x\|$ denotes the Euclidean norm of x. The set of real symmetric matrices of size $d \times d$ is denoted $\mathcal{S}^d$, and its subsets of positive semi-definite matrices and of positive definite matrices are denoted $\mathcal{S}^d_+$ and $\mathcal{S}^d_{++}$, respectively. Given a positive integer n, $\mathfrak{S}_n$ represents the set of permutations of $\{1, \dots, n\}$, and its cardinality is denoted $n!$. The differential application of a function F at a point x is denoted $\mathcal{D}F(x)$, and the Clarke derivative of F at x is denoted $\partial F(x)$. We use the notations $\operatorname{Arg\,min} f$ and $\operatorname{Arg\,max} f$ for the sets of global minima and global maxima of f, respectively. When unique global minimum and maximum exist, we denote them as $\arg\min f$ and $\arg\max f$, respectively. For any sequence $\{v_k\}_{k \in \mathbb{N}^*}$ and any $k \in \mathbb{N}^*$, we set $v_{1:k} = (v_1, \dots, v_k)$. For a topological space X, we denote $\mathcal{B}(\mathsf{X})$ the Borel σ-field of X, which makes X a measured space. If μ is a measure on $\mathcal{B}(\mathsf{X})$ and ν is a measure on $\mathcal{B}(\mathsf{Y})$, we denote $\mu \otimes \nu$ the product measure of μ and ν, which is a measure on $\mathcal{B}(\mathsf{X} \times \mathsf{Y})$. Likewise, for $k \in \mathbb{N}^*$, we denote $\mu^{\otimes k}$ the measure product of μ by itself k times, as a measure on $\mathcal{B}(\mathsf{X}^k)$.

2 Definition of Markov chains arising from a normalization of CMA-ES

We present in this section the CMA-ES algorithm and define normalized Markov chains—candidates to be stable—associated to the algorithm. We explain the connection between the stability of these Markov chains and the convergence of the algorithm, motivating thus why the irreducibility, aperiodicity and topological properties of the Markov chains that we study in the paper are an important part for obtaining a convergence proof of CMA-ES.

2.1 Presentation of CMA-ES

The covariance matrix adaptation evolution strategy (CMA-ES) is an iterative algorithm which aims to approximate a problem solution

$$x^* \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^d} f(x) \tag{P}$$

where $d \in \mathbb{N}^*$ is the dimension of the problem, and $f \colon \mathbb{R}^d \to \mathbb{R}$ is the objective function. A vector $x^* \in \mathbb{R}^d$, solution to (P), is called a global minimum of f. The CMA-ES attempts to approach x^* by successively sampling,

for iterations $t \in \mathbb{N}$, new candidate solutions from a multivariate normal probability distribution $\mathcal{N}(m_t, \sigma_t^2 \mathbf{C}_t)$. The vector $m_t \in \mathbb{R}^d$ is the current mean of the distribution and we specifically desire that $f(m_t)$ converges to the essential infimum of f. The positive real number $\sigma_t > 0$ is the current stepsize, and the symmetric positive definite matrix $\mathbf{C}_t \in \mathcal{S}_{++}^d$ is referred to as the current covariance matrix. For our analysis, we generalize the assumption that the distribution of the candidate solutions is multivariate normal.

The parameters of the sampling distribution are updated using two cumulation paths $p_t^\sigma, p_t^c \in \mathbb{R}^d$, which implement a weighted moving average of the steps followed by the mean.

More precisely, the algorithm works as follows. At iteration $t \in \mathbb{N}$, given $m_t \in \mathbb{R}^d$, $\sigma_t > 0$, $\mathbf{C}_t \in \mathcal{S}_{++}^d$, and $p_t^\sigma, p_t^c \in \mathbb{R}^d$, we generate independent identically distributed (i.i.d.) samples $U_{t+1}^1, \ldots, U_{t+1}^\lambda$ following a sampling distribution ν_U^d in $\mathbb{R}^d$ and independently of $(m_t, \sigma_t, \mathbf{C}_t, p_t^\sigma, p_t^c)$. Usually, the distribution ν_U^d is the standard normal distribution in $\mathbb{R}^d$. However, throughout the paper we will refer to CMA-ES as the algorithm presented in this section with a general and abstract sampling distribution ν_U^d. We compute then λ candidate solutions

$$x_{t+1}^i \coloneqq m_t + \sigma_t \sqrt{\mathbf{C}_t} U_{t+1}^i \text{ for } i = 1, \ldots, \lambda \ , \tag{2.1}$$

and rank them with respect to their f-values. Formally, we define a permutation $s_{t+1} \in \mathfrak{S}_\lambda$ satisfying

$$f\left(x_{t+1}^{s_{t+1}(1)}\right) \leqslant \cdots \leqslant f\left(x_{t+1}^{s_{t+1}(\lambda)}\right) \ . \tag{2.2}$$

When $f\left(x_{t+1}^i\right) = f\left(x_{t+1}^j\right)$, we impose for uniqueness $s_{t+1}^{-1}(i) < s_{t+1}^{-1}(j)$ if $i < j$. We say that we have *neutral selection* when, instead of (2.2), the permutation s_{t+1} is independent of the samples U_{t+1}^i for all $t \in \mathbb{N}$. This is the case, for example, when the permutation is fixed for all t, or when $f(x_{t+1}^i)$ is independent of its argument or independent of U_{t+1}^i.

The mean is moved towards the best solutions, and is updated by applying the function $F_{c_m}^{\mathrm{m}}$ defined as

$$F_{c_m}^{\mathrm{m}} \colon (m, v) \in \mathbb{R}^d \times \mathbb{R}^d \quad \mapsto m + c_m v \ , \tag{2.3}$$

given a fixed learning rate $c_m > 0$ (by default $c_m = 1$). Precisely, the mean obeys

$$m_{t+1} = F_{c_m}^{\mathrm{m}}\left(m_t, \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}\right) = m_t + c_m \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)} \ , \tag{2.4}$$

where $w_1^m \geqslant \cdots \geqslant w_\mu^m > 0$ are weights such that $\sum_{i=1}^\mu w_i^m = 1$, and $\sqrt{\mathbf{C}_t}$ is the symmetric positive definite square root of $\mathbf{C}_t$.

We introduce the function to update the paths $p_t^\sigma, p_t^c \in \mathbb{R}^d$. Given a decay factor $c \in (0, 1]$, F_c^{p} is defined as

$$F_c^{\mathrm{p}} \colon (p, v) \in \mathbb{R}^d \times \mathbb{R}^d \quad \mapsto (1 - c)p + \sqrt{c(2 - c)\mu_{\mathrm{eff}}}\, v \tag{2.5}$$

where $\mu_{\mathrm{eff}} = 1/\|\mathbf{w}_m\|^2$, with $\mathbf{w}_m = (w_1^m, \ldots, w_\mu^m)^\top$. The closer the decay factor c is to zero, the more the updated path depends on the previous path due to the term $(1 - c)p$. Conversely, when $c = 1$, the updated path is collinear to and only depends on v. We set two decay factors, $c_\sigma, c_c \in (0, 1]$, and use (2.5) to update two cumulation paths, one for updating the stepsize and the other for the rank-one update of the covariance matrix (see below). We update

$$p_{t+1}^\sigma = (1 - c_\sigma)p_t^\sigma + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\mathrm{eff}}} \sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)} = F_{c_\sigma}^{\mathrm{p}}\left(p_t^\sigma, \sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}\right) \ , \tag{2.6}$$

and

$$p_{t+1}^c = (1 - c_c)p_t^c + \sqrt{c_c(2 - c_c)\mu_{\mathrm{eff}}} \sqrt{\mathbf{C}_t} \sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)} = F_{c_c}^{\mathrm{p}}\left(p_t^c, \sqrt{\mathbf{C}_t} \sum_{i=1}^\mu w_i^m U_{t+1}^{s_{t+1}(i)}\right) \ . \tag{2.7}$$

The second argument in the RHS of (2.7) is the same as in (2.4) disregarding stepsize σ_t. Eq. (2.6) additionally drops $\sqrt{\mathbf{C}_t}$. Consequently, when p_0^σ and U_{t+1}^i are standard Gaussian, then, under neutral selection, p_{t+1}^σ is a standard Gaussian vector too and, in particular, the length of p_{t+1}^σ does not depend on its direction. The path p_{t+1}^c from (2.7) maintains under neutral selection the covariance matrix $\mathbf{C}_t$ when p_t^c has covariance matrix $\mathbf{C}_t$. The path is commensurable with updating $\mathbf{C}_t$ and its expected length can strongly depend on its direction.

The stepsize is updated using the path p_{t+1}^σ. Considering an abstract measurable function $\Gamma \colon \mathbb{R}^d \to \mathbb{R}_{++}$ that we call stepsize change, the update reads

$$\sigma_{t+1} = \sigma_t \times \Gamma\left(p_{t+1}^\sigma\right) \ . \tag{2.8}$$

A standard stepsize change used in CMA-ES is the cumulative stepsize adaptation (CSA) where Γ equals

$$\Gamma^1_{\mathrm{CSA}}(p) = \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p\|}{\mathbb{E}\|\nu_U^d\|} - 1\right)\right) \ , \tag{2.9}$$

where $\mathbb{E}\|\nu_U^d\| := \mathbb{E}\|\xi\|$ for a random variable ξ distributed under ν_U^d. When ν_U^d is the standard normal distribution, (2.9) increases the stepsize when $\|p_t^\sigma\|$ is larger than to be expected under neutral selection (assuming that $p_0^\sigma \sim \nu_U^d$) and decreases the stepsize when $\|p_t^\sigma\|$ is smaller. When consecutive steps are taken in a similar direction, the expected path is long while the same progress could be made in fewer iterations with larger steps. When consecutive steps are negatively correlated, the expected path is short and a smaller stepsize is advisable. A smooth alternative to (2.9) implementing the same idea is [4]

$$\Gamma^2_{\mathrm{CSA}}(p) = \exp\left(\frac{c_\sigma}{2d_\sigma}\left(\frac{\|p\|^2}{\mathbb{E}\|\nu_U^d\|^2} - 1\right)\right) \ . \tag{2.10}$$

These two stepsize changes rely on the choice of damping parameter $d_\sigma > 0$, which is chosen $\approx 1 + 2\sqrt{\mu_{\mathrm{eff}}/d}$ in the first case and $\approx 1 + 2\mu_{\mathrm{eff}}/d$ in the second case. Empirically, Γ^1_{CSA} and Γ^2_{CSA} show similar performance when used with CMA-ES [18]. While the function Γ^1_{CSA} is the default stepsize change, previous theoretical works on ES also have investigated Γ^2_{CSA} [4, 42].

Last, we introduce the update function for the covariance matrix, which depends on the choice of learning rates $c_1, c_\mu \geqslant 0$ such that $c_1 + c_\mu \in [0, 1]$:

$$\begin{aligned} F^{\mathrm{C}}_{c_1,c_\mu} \colon \mathcal{S}^d_{++} \times \mathbb{R}^d \times \mathcal{S}^d_{+} &\to \mathcal{S}^d_{++} \\ (\mathbf{C}, p, \mathbf{M}) &\mapsto (1 - c_1 - c_\mu)\mathbf{C} + c_1 pp^\top + c_\mu \mathbf{M} \ , \end{aligned} \tag{2.11}$$

and the covariance matrix is updated via

$$\begin{aligned} \mathbf{C}_{t+1} &= (1 - c_1 - c_\mu)\mathbf{C}_t + c_1 p^{\mathrm{c}}_{t+1}[p^{\mathrm{c}}_{t+1}]^\top + c_\mu \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^{\mathrm{c}} \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{C}_t} \\ &= F^{\mathrm{C}}_{c_1,c_\mu}\left(\mathbf{C}_t, p^{\mathrm{c}}_{t+1}, \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^{\mathrm{c}} \left[U_{t+1}^{s_{t+1}(i)}\right] \left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\mathbf{C}_t}\right) \ , \end{aligned} \tag{2.12}$$

where we define weights $w_1^{\mathrm{c}} \geqslant \cdots \geqslant w_\mu^{\mathrm{c}} > 0$ such that $\sum_{i=1}^{\mu} w_i^{\mathrm{c}} = 1$. Moreover, we assume throughout the paper that $0 < c_1 + c_\mu < 1$. This assumption will be essential in the proofs of Lemma 5.6, Corollary 5.1 and Proposition 5.7. The setting $c_1 + c_\mu = 1$ is however used in practice when μ is large and we believe that with further work our results could be proven for this case as well.

The term $c_1 pp^\top$ is called the rank-one update, whereas the term $c_\mu \mathbf{M}$ is the rank-mu update since in (2.12) we replace $\mathbf{M}$ by a matrix of rank $\min(\mu, d)$ almost surely which satisfies a maximum likelihood condition for the best samples of the last iteration [22, Proposition 7]. In practice, also negative weights are used for the rank-mu update of the covariance matrix [30, 27]. However, since the updated covariance matrix must be positive definite, the norm of the vectors corresponding to negative weights must be controlled. We do not consider negative weights in the present paper.

2.2 Assumptions

Our analysis of CMA-ES relies on analyzing the stability of normalized Markov chains underlying the algorithm. The construction of these Markov chains assumes that the objective function is scaling-invariant. A function $f \colon \mathbb{R}^d \to \mathbb{R}$ is said to be *scaling-invariant with respect to* $x^* \in \mathbb{R}^d$ when for all $x, y \in \mathbb{R}^d$ and $\rho > 0$:

$$f(x + x^*) \leqslant f(y + x^*) \Leftrightarrow f(\rho x + x^*) \leqslant f(\rho y + x^*) \ . \tag{2.13}$$

The class of scaling-invariant functions has been of interest for the convergence analysis of different variants of ES [8, 42], and is related to the class of positively homogeneous functions [43]. We make an additional technical assumption on the level sets of the objective function to avoid ties in (2.2), which will be useful to define lower semi-continuous density functions in Lemma 5.2. Overall, we will use the following assumptions:

F1. *The objective function $f \colon \mathbb{R}^d \to \mathbb{R}$ is a strictly increasing transformation of a continuous function with Lebesgue negligible level sets.*

F2. *The objective function* $f\colon \mathbb{R}^d \to \mathbb{R}$ *is scaling-invariant with respect to a point* $x^* \in \mathbb{R}^d$.

Instead of assuming **F**1, we can suppose without loss of generality that the function f is continuous with Lebesgue negligible level sets since CMA-ES is invariant to increasing transformations of the objective function [6]. Assumption **F**2 is central in this analysis since it is required to define an equivalent, normalized Markov chain via (2.14) below.

In order to go beyond scaling-invariant functions, it might be possible to adopt another approach, considering for instance recent works on the convergence of evolution strategies that prove a drift condition on the state variables of the algorithm and hence the convergence on composites of strongly convex functions with strictly increasing functions (for however so far the (1+1)-ES selection scheme only) [3, 36].

Furthermore, the sampling distribution ν_U^d should satisfy the following assumption that allows in particular to characterize a density for the ranked candidate solutions, see Lemma 5.2.

N1. *The probability distribution* ν_U^d *admits a continuous density* $p_U^d(\cdot)$ *with respect to the Lebesgue measure on* $\mathbb{R}^d$ *which is positive everywhere on* $\mathbb{R}^d$.

This assumption is satisfied by a multivariate standard normal distribution as used in CMA-ES. We have moreover the following assumptions on the stepsize change Γ.

Γ1. *The stepsize change* $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ *is a locally Lipschitz map and is differentiable at every nonzero vector of* $\mathbb{R}^d$.

Γ2. *Given* c_c *the cumulation parameter for the path in* (2.6)*, the function* Γ *satisfies* $\liminf \Gamma(p) > (1-c_c)^{-1}$ *for* $\|p\|$ *to* $+\infty$.

Γ3. *The function* Γ *satisfies* $\Gamma(0) < 1$.

Assumption **Γ**1 is required to apply the results stated in Section 4.1 and in particular to ensure that the analyzed process satisfies the condition **H**2 in Section 4.1 to obtain the irreducibility and aperiodicity of the Markov chain defined in (2.14). Assumptions **Γ**2 and **Γ**3 are used in Propositions 5.5 and 5.7, respectively.

Assumptions **Γ**1–**Γ**3 are satisfied by both stepsize changes, Γ^1_{CSA} and Γ^2_{CSA}, as stated in the following lemma.

Lemma 2.1. *Assume that* $c_\sigma \in (0,1]$. *Then, the stepsize change functions* Γ^1_{CSA} *and* Γ^2_{CSA}, *defined by* (2.9) *and* (2.10) *respectively, satisfy the assumptions* **Γ**1–**Γ**3.

Proof. The proof is simple and left to the reader. □

2.3 Proving the stability of a normalized Markov chain leads to linear convergence

Before stating our main results, we define a normalized Markov chain underlying the CMA-ES algorithm. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event is an element $\mathsf{W} \in \mathcal{F}$, and the probability of W is $\mathbb{P}[\mathsf{W}]$. A random variable U valued in a measured space $(\mathsf{U}, \mathcal{U})$ is defined as a measurable function $U\colon \Omega \to U$, and for $\mathsf{A} \in \mathcal{U}$, we identify $\mathbb{P}[U \in \mathsf{A}]$ to $\mathbb{P}[\{\omega \in \Omega \mid U(\omega) \in \mathsf{A}\}]$. A transition kernel on a topological state space X equipped with its Borelian σ-field $\mathcal{B}(\mathsf{X})$ is an application $P\colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}_+$ such that, for every $x \in \mathsf{X}$, $\mathsf{A} \in \mathcal{B}(\mathsf{X}) \mapsto P(x, \mathsf{A})$ is a probability measure, and for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, $x \in \mathsf{X} \mapsto P(x, \mathsf{A})$ is a measurable map. Then, a (time-homogeneous) Markov chain with transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and initial probability distribution ν on $\mathcal{B}(\mathsf{X})$ is a sequence of random variables $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ valued in X, satisfying for every $t \in \mathbb{N}$

$$\mathbb{P}\left[(\phi_0, \ldots, \phi_t) \in \mathsf{A}_0 \times \cdots \times \mathsf{A}_t \mid \phi_0 \sim \nu\right] = \int_{\mathsf{X}^{t+1}} \mathbb{1}\left\{(x_0, \ldots, x_{t-1}) \in \mathsf{A}_0 \times \cdots \times \mathsf{A}_{t-1}\right\} P(x_{t-1}, \mathsf{A}_t) P(x_{t-2}, \mathrm{d}x_{t-1}) \ldots P(x_0, \mathrm{d}x_1) \nu(\mathrm{d}x_0)$$

where for every probability measure ν on $\mathcal{B}(\mathsf{X})$, we have equipped $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}[\cdot|\phi_0 \sim \nu]$. We define the t-step transition kernel by $P^t(x, \mathsf{A}) = \mathbb{P}[\phi_t \in \mathsf{A} | \phi_0 \sim \delta_x]$ for every $t \geqslant 0$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$.

The sequence $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ introduced in Section 2.1 defines a time-homogeneous Markov chain on the state space $\mathbb{R}^{3d} \times \mathbb{R}_{++} \times \mathcal{S}^d_{++}$. This is immediate from the observation that the definition of $(m_{t+1}, p_{t+1}^\sigma, p_{t+1}^c, \sigma_{t+1}, \mathbf{C}_{t+1})$ depends only on the previous state $(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)$ and the independent random input $U_{t+1}^1, \ldots, U_{t+1}^\lambda$. However, when the mean converges to the optimum of the function, the stepsize σ_t, the covariance matrix $\mathbf{C}_t$ and the path p_t^c converge to 0. Therefore, this Markov chain is not Harris recurrent (it does not revisit every neighborhood of any state infinitely many times). Yet, as illustrated later in (2.18) and Proposition 2.3, our methodology to prove

linear convergence [8] relies on a Law of Large Numbers which motivates to have a positive Harris recurrent Markov chain (with a stationary probability distribution) and, more generally, a chain stable enough to apply an ergodic theorem [35, Theorems 13.0.1] and satisfy a Law of Large Numbers [35, Theorem 17.0.1]. Therefore, we define a normalized process, candidate to have a stationary distribution, underlying the CMA-ES algorithm. Consider $R\colon \mathcal{S}_{++}^d \to \mathbb{R}_{++}$ a normalization function which is

R1. *(positively) homogeneous with degree 1, i.e., for every* $\mathbf{A} \in \mathcal{S}_{++}^d$ *and* $\rho > 0, R(\rho\mathbf{A}) = \rho R(\mathbf{A})$,

R2. *locally Lipschitz continuous,*

R3. *differentiable on a nonempty open subset of* $\mathcal{S}_{++}^d$.

Assumption **R**1 is required to define a normalized Markov chain, see (2.14) below, as proven in Proposition 2.2. Assumption **R**2 is used to prove irreducibility and aperiodicity of the normalized chain, notably for the verification of condition **H**2 introduced in Section 4.1. Later, Proposition 5.7 uses **R**3 to prove a maximal rank condition.

We give examples of normalization functions that satisfy these assumptions.

Proposition 2.1. *The* d-th *root of the determinant,* $\det(\cdot)^{1/d}$, *and the* i-th *largest eigenvalue,* $\lambda_i(\cdot)$, *for* $i \in \{1, \dots, d\}$ *counted with multiplicity, are functions defined on* $\mathcal{S}_{++}^d$ *that satisfy* **R**1–**R**3.

Proof. The proof of the property **R**1 is immediate from the linearity of the determinant as a function of the columns of the matrix and the definition of an eigenvalue. For the properties **R**2 and **R**3, we know that the determinant of a matrix is a polynomial function of the coefficients of the matrix [29, Section 0.3], hence it is infinitely differentiable and in particular is locally Lipschitz [13, Proposition and Corollary 2.2.1]. For the eigenvalues, $\lambda_i(\cdot)$ is locally Lipschitz on $\mathcal{S}_{++}^d$, as a consequence of Weyl's theorem [29, Corollary 4.3.15]. Besides, $\lambda_i(\cdot)$ is infinitely differentiable on a neighborhood of any symmetric matrix with eigenvalues that have simple multiplicity [40, Theorem 5.3]. □

Given the CMA-ES Markov chain $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$ defined in Section 2.1 and a normalization function R, we define the *normalized chain* $\Phi = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ as follows.[2] For all $t \geqslant 1$, we set

$$z_t = \frac{m_t - x^*}{\sigma_t \sqrt{R(\mathbf{C}_t)}},\ p_t = p_t^\sigma,\ q_t = \frac{p_t^c}{\sqrt{R(\mathbf{C}_{t-1})}},\ \mathbf{\Sigma}_t = \frac{\mathbf{C}_t}{R(\mathbf{C}_t)},\ r_t = \frac{R(\mathbf{C}_t)}{R(\mathbf{C}_{t-1})}\ . \tag{2.14}$$

We prove below that when the objective function is scaling-invariant, the normalized chain defined by (2.14) is a time-homogeneous Markov chain that can be defined independently of the original Markov chain $\{(m_t, p_t^\sigma, p_t^c, \sigma_t, \mathbf{C}_t)\}_{t\in\mathbb{N}}$. We establish first that on scaling-invariant functions, the permutation sorting the candidate solutions $m_t + \sigma_t\sqrt{\mathbf{C}_t}U_{t+1}^i$ also sorts the vectors $z_t + \sqrt{\mathbf{\Sigma}_t}U_{t+1}^i$, for $i = 1, \dots, \lambda$.

Lemma 2.2. *Let* $t \geqslant 1$ *and suppose that the objective function* f *satisfies* **F**2. *Let* $s_{t+1} \in \mathfrak{S}_\lambda$ *be a (random) permutation that sorts the indices* $i = 1, \dots, \lambda$ *with respect to the* f*-values of* $m_t + \sigma_t\sqrt{\mathbf{C}_t}U_{t+1}^i$. *Then,* s_{t+1} *also sorts the indices* $i = 1, \dots, \lambda$ *with respect to the* f*-values of* $x^* + z_t + \sqrt{\mathbf{\Sigma}_t}U_{t+1}^i$. *Moreover, we can ensure the uniqueness of* s_{t+1} *by imposing a tie-break (cf. Section 2.1).*

Proof. Let $i = 1, \dots, \lambda$. By definition of z_t and $\mathbf{\Sigma}_t$, we obtain

$$f\left(x^* + z_t + \sqrt{\mathbf{\Sigma}_t}U_{t+1}^i\right) = f\left(x^* + R(\mathbf{C}_t)^{-1/2}\sigma_t^{-1} \times \left[m_t - x^* + \sigma_t\sqrt{\mathbf{C}_t}U_{t+1}^i\right]\right).$$

We conclude by using the definition of a scaling-invariant function (2.13). □

From the previous lemma, we deduce in Proposition 2.2 below that the normalized chain defined in (2.14) is a time-homogeneous Markov chain that can be defined independently of the original Markov chain. Indeed, given R satisfying **R**1, denote with a slight abuse of notation (since we use the same notation as for (2.14) with however a different time index) the time-homogeneous Markov chain $\Phi = \{\phi_t\}_{t\geqslant 0} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 0}$ defined via

[2] The definition of q_t in (2.14) suggests transforming p_t^c in (2.7) like $\mathbf{C}_t^{1/2}\mathbf{C}_{t-1}^{-1/2}p_t^c$ to avoid the time index $t-1$ in (2.14). Then, p_{t+1}^c would become equal to $\mathbf{C}_t^{1/2}p_{t+1}^\sigma$. We can prove unbiasedness for p^σ [22] and affine invariance with p^c and $c_\sigma = 1$ [6].

$\phi_0 \in \mathsf{Y} = (\mathbb{R}^d)^3 \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ (where $R^{-1}(\{1\}) = \{\boldsymbol{\Sigma} \in \mathcal{S}_{++}^d : R(\boldsymbol{\Sigma}) = 1\}$) and the following recursion

$$
\begin{aligned}
z_{t+1} &= \frac{z_t + c_m \sqrt{\boldsymbol{\Sigma}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}}}{\sqrt{r_{t+1}}\,\Gamma(p_{t+1})} = \frac{F_{c_m}^{\mathrm{m}}(z_t, \sqrt{\boldsymbol{\Sigma}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}})}{\sqrt{r_{t+1}}\,\Gamma(p_{t+1})} \\
p_{t+1} &= (1-c_\sigma)p_t + \sqrt{c_\sigma(2-c_\sigma)\mu_{\mathrm{eff}}}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}} = F_{c_\sigma}^{\mathrm{p}}(p_t, \mathbf{w}_m^\top U_{t+1}^{s_{t+1}}) \\
q_{t+1} &= F_{c_c}^{\mathrm{p}}(r_t^{-1/2} q_t, \sqrt{\boldsymbol{\Sigma}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}}) \\
\boldsymbol{\Sigma}_{t+1} &= r_{t+1}^{-1} F_{c_1,c_\mu}^{\mathrm{C}}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_t}\right) \\
r_{t+1} &= R \circ F_{c_1,c_\mu}^{\mathrm{C}}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_t}\right)
\end{aligned}
\tag{2.15}
$$

with $\mathbf{U} = \{U_{t+1}\}_{t\in\mathbb{N}}$ an i.i.d. process independent of ϕ_0 with $U_1 = (U_1^1, \dots, U_1^\lambda) \sim (\nu_U^d)^{\otimes\lambda}$, and s_{t+1} the (almost surely unique) permutation that sorts the $f(z_t + \sqrt{\boldsymbol{\Sigma}_t}U_{t+1}^i)$, $i = 1, \dots, \lambda$. Moreover, $U_{t+1}^{s_{t+1}}$ denotes the collection of vectors $(U_{t+1}^{s_{t+1}(1)}, \dots, U_{t+1}^{s_{t+1}(\lambda)})$. Remark that in (2.15), the update of the covariance matrix $\boldsymbol{\Sigma}_{t+1}$ writes

$$
\boldsymbol{\Sigma}_{t+1} = \frac{\tilde{\boldsymbol{\Sigma}}_{t+1}}{R(\tilde{\boldsymbol{\Sigma}}_{t+1})}
$$

where $\tilde{\boldsymbol{\Sigma}}_{t+1}$ is the covariance matrix to which we apply the rank-one and rank-mu updates, i.e.,

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_{t+1} &\coloneqq F_{c_1,c_\mu}^{\mathrm{C}}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_t}\right) \\
&= (1-c_1-c_\mu)\boldsymbol{\Sigma}_t + c_1 q_{t+1}(q_{t+1})^\top + c_\mu \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top \sqrt{\boldsymbol{\Sigma}_t}\ ,
\end{aligned}
\tag{2.16}
$$

where $F_{c_1,c_\mu}^{\mathrm{C}}$ is defined via (2.11). Similarly r_{t+1} can be expressed using $\tilde{\boldsymbol{\Sigma}}_{t+1}$ as

$$
r_{t+1} = R(\tilde{\boldsymbol{\Sigma}}_{t+1})\ .
$$

The update of ϕ_t in (2.15) defines a function F_Φ such that

$$
\phi_{t+1} = F_\Phi(\phi_t, U_{t+1}^{s_{t+1}})\ . \tag{2.17}
$$

We prove in the next proposition that if f is scaling-invariant, the normalized chain defined in (2.14) can be defined independently of the original CMA-ES chain via the recursion (2.17) provided it is initialized properly. While the normalized process (2.14) imposes $t \geqslant 1$, the next proposition defines this process via the recursion (2.15) and thus allows to start with any time index.

Proposition 2.2. *Suppose that the objective function f satisfies* **F**2 *and that the normalization function R satisfies* **R**1. *Let $\{(m_t, p_t^\sigma, p_t^c, \mathbf{C}_t, \sigma_t)\}_{t\in\mathbb{N}}$ be the chain associated to CMA-ES defined in Section 2.1 and $\Phi = \{\phi_t\}_{t\geqslant 1} = \{(z_t, p_t, q_t, \boldsymbol{\Sigma}_t, r_t)\}_{t\geqslant 1}$ be the normalized process defined via (2.14) for $t \geqslant 1$. Then Φ is a time-homogeneous Markov chain valued in the state space $\mathsf{Y} = (\mathbb{R}^d)^3 \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ that satisfies*

$$
\phi_1 = \left(\frac{m_1}{\sigma_1\sqrt{R(\mathbf{C}_1)}}, p_1^\sigma, \frac{p_1^c}{\sqrt{R(\mathbf{C}_0)}}, \frac{\mathbf{C}_1}{R(\mathbf{C}_1)}, \frac{R(\mathbf{C}_1)}{R(\mathbf{C}_0)}\right)
$$

and for $t \geqslant 1$ we have

$$
\phi_{t+1} = F_\Phi(\phi_t, U_{t+1}^{s_{t+1}})
$$

where F_Φ is the function in (2.17) defined via the equations (2.15), $s_{t+1} \in \mathfrak{S}_\lambda$ is a permutation that sorts[3] the $f(x^* + z_t + \sqrt{\boldsymbol{\Sigma}_t}U_{t+1}^i)$, $i = 1, \dots, \lambda$, and $\mathbf{U} = \{U_{t+1}\}_{t\geqslant 1}$ is the i.i.d. process used to define $\{(m_t, p_t^\sigma, p_t^c, \mathbf{C}_t, \sigma_t)\}_{t\in\mathbb{N}}$, thus independent of ϕ_1.

[3] We always sort increasing and, as explained in Section 2.1, in case of a tie between the f-values of the candidate solutions of indices i and j with $i < j$, we impose $s_{t+1}^{-1}(i) < s_{t+1}^{-1}(j)$ to ensure the uniqueness of the permutation s_{t+1}.

Proof. By Lemma 2.2, it is sufficient to show that (2.15) holds for every $t \geqslant 1$ in order to prove that Φ is a time-homogeneous Markov chain. Let $t \geqslant 1$, and consider the matrix $\tilde{\mathbf{\Sigma}}_{t+1}$ defined in (2.16). Since $F^{\mathrm{C}}_{c_1,c_\mu}$ is homogeneous with respect to its first variable, positively homogeneous of degree 2 with respect to the second variable, using (2.14) and the definition of $\mathbf{C}_{t+1}$ in (2.12) we find

$$\tilde{\mathbf{\Sigma}}_{t+1} = R(\mathbf{C}_t)^{-1}\mathbf{C}_{t+1} \ .$$

By the property **R**1 applied to the previous equation we obtain $R(\tilde{\mathbf{\Sigma}}_{t+1}) = R(\mathbf{C}_t)^{-1}R(\mathbf{C}_{t+1}) = r_{t+1}$. Furthermore, the following holds

$$\begin{aligned}
z_{t+1} =& R(\mathbf{C}_{t+1})^{-1/2}\sigma_{t+1}^{-1} \times (m_{t+1} - x^*) \\
=& r_{t+1}^{-1/2} R(\mathbf{C}_t)^{-1/2}\sigma_t^{-1}\Gamma(p^\sigma_{t+1})^{-1} \times \left[m_t - x^* + c_m \sigma_t \sqrt{\mathbf{C}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right] \\
=& r_{t+1}^{-1/2}\Gamma(p_{t+1})^{-1} \times \left[z_t + c_m \sqrt{\mathbf{\Sigma}_t} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \right] = \frac{F^{\mathrm{m}}_{c_m}(z_t, \sqrt{\mathbf{\Sigma}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}})}{\sqrt{r_{t+1}}\Gamma(p_{t+1})} \ ,
\end{aligned}$$

where $F^{\mathrm{m}}_{c_m}$ is defined via (2.3). Moreover,

$$\begin{aligned}
\mathbf{\Sigma}_{t+1} &= R(\mathbf{C}_{t+1})^{-1}\mathbf{C}_{t+1} \\
&= R(\mathbf{C}_{t+1})^{-1} \left[(1 - c_1 - c_\mu)\mathbf{C}_t + c_1 (p^c_{t+1})(p^c_{t+1})^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left(\sqrt{\mathbf{C}_t} U_{t+1}^{s_{t+1}(i)} \right) \left(\sqrt{\mathbf{C}_t} U_{t+1}^{s_{t+1}(i)} \right)^\top \right] \\
&= r_{t+1}^{-1} \times \left[(1 - c_1 - c_\mu)\mathbf{\Sigma}_t + c_1 (q_{t+1})(q_{t+1})^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \left(\sqrt{\mathbf{\Sigma}_t} U_{t+1}^{s_{t+1}(i)} \right) \left(\sqrt{\mathbf{\Sigma}_t} U_{t+1}^{s_{t+1}(i)} \right)^\top \right] \\
&= r_{t+1}^{-1} F^{\mathrm{C}}_{c_1,c_\mu} \left(\mathbf{\Sigma}_t, q_{t+1}, \sqrt{\mathbf{\Sigma}_t} \sum_{i=1}^{\mu} w_i^c \left[U_{t+1}^{s_{t+1}(i)} \right] \left[U_{t+1}^{s_{t+1}(i)} \right]^\top \sqrt{\mathbf{\Sigma}_t} \right)
\end{aligned}$$

Finally,

$$\begin{aligned}
q_{t+1} &= R(\mathbf{C}_t)^{-1/2} p^c_{t+1} \\
&= R(\mathbf{C}_t)^{-1/2}(1 - c_c)p^c_t + \sqrt{\mu_{\mathrm{eff}} c_c (2 - c_c)} R(\mathbf{C}_t)^{-1/2} \mathbf{C}_t^{1/2} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} \\
&= r_t^{-1/2}(1 - c_c)q_t + \sqrt{\mu_{\mathrm{eff}} c_c (2 - c_c)} \mathbf{\Sigma}_t^{1/2} \sum_{i=1}^{\mu} w_i^m U_{t+1}^{s_{t+1}(i)} = F^{\mathrm{p}}_{c_c}(r_t^{-1/2} q_t, \sqrt{\mathbf{\Sigma}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}}) \ ,
\end{aligned}$$

where $F^{\mathrm{p}}_{c_c}$ is defined via (2.5). □

Now that we formally prove that the normalized chain defined in (2.14) is a time-homogeneous Markov chain when the algorithm optimizes a scaling-invariant function, we recapitulate how its stability is connected to the linear convergence of CMA-ES on scaling-invariant functions. For $T \in \mathbb{N}$, using the definition of the normalized Markov chain in (2.14) and the definition of the stepsize change (2.8) we obtain

$$\begin{aligned}
\frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} &= \frac{1}{T} \sum_{t=0}^{T-1} [\log \|m_{t+1} - x^*\| - \log \|m_t - x^*\|] \\
&= \frac{1}{T} \sum_{t=0}^{T-1} \left(\log \left(\|z_{t+1}\| \sqrt{R(\mathbf{C}_{t+1})} \sigma_{t+1} \right) - \log \left(\|z_t\| \sqrt{R(\mathbf{C}_t)} \sigma_t \right) \right) \\
&= \frac{1}{T} \sum_{t=0}^{T-1} \left(\log \|z_{t+1}\| - \log \|z_t\| + \log \frac{\sigma_{t+1}}{\sigma_t} + \frac{1}{2} \log \frac{R(\mathbf{C}_{t+1})}{R(\mathbf{C}_t)} \right) \\
&= \frac{1}{T} \sum_{t=0}^{T-1} \left(\log \|z_{t+1}\| - \log \|z_t\| + \log \Gamma(p_{t+1}) + \frac{1}{2} \log r_{t+1} \right) \ . \qquad (2.18)
\end{aligned}$$

If the Law of Large Numbers applies to the RHS of (2.18), we obtain a limit of the LHS when T goes to infinity. If this limit is proven to be strictly negative, we have shown linear convergence of the underlying optimization algorithm. In order to apply limit theorems [35, Theorem 17.0.1] and obtain a Law of Large Numbers, we require the chain Φ to be geometrically ergodic. Key assumptions for ergodicity are irreducibility and aperiodicity of the Markov chain whose notions will be formally introduced in Section 3. We thus connected the stability of the normalized chain to the convergence of the underlying optimization algorithm. More formally the following proposition holds.

Proposition 2.3. *Consider the CMA-ES algorithm defined in Section 2.1 optimizing a function f satisfying* **F**2. *Assume that the process Φ, obeying (2.15) with state space $\mathsf{Y} = (\mathbb{R}^d)^3 \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ (where $R^{-1}(\{1\}) = \{\mathbf{\Sigma} \in \mathcal{S}^d_{++} : R(\mathbf{\Sigma}) = 1\}$), is an irreducible, aperiodic and positive Harris-recurrent Markov chain with (unique) invariant probability measure π. Assume moreover that the functions*

$$(z, p, q, \mathbf{\Sigma}, r) \in \mathsf{Y} \mapsto \log \|z\|, \log \Gamma(p), \log r \tag{2.19}$$

are π-integrable. Then the CMA-ES algorithm behaves globally asymptotically linearly almost surely:

$$\lim_{T\to\infty} \frac{1}{T} \log \frac{\|m_T - x^*\|}{\|m_0 - x^*\|} = \lim_{t\to\infty} \mathbb{E}\left[\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|}\right] = \int \left(\log \Gamma(p) + \frac{1}{2} \log r\right) \mathrm{d}\pi \ . \tag{2.20}$$

Proof. The almost sure limit of the LHS in (2.20) follows directly from (2.18), LLN for ergodic chains [35, Theorem 17.0.1]. The limit of the expectation in (2.20) follows from the ergodic theorem [35, Theorem 14.0.1], since

$$\log \frac{\|m_{t+1} - x^*\|}{\|m_t - x^*\|} = \log \|z_{t+1}\| - \log \|z_t\| + \log \Gamma(p_{t+1}) + \frac{1}{2} \log r_{t+1} \ .$$

□

The previous proposition illustrates that proving irreducibility and aperiodicity of the chain Φ is a stepping stone to establish linear convergence of CMA-ES. Proving these properties will occupy Section 5. Later, we intend to prove the geometric ergodicity by means of Foster-Lyapunov drift conditions [35, Theorem 15.0.1] that depend on small sets (as given in Theorem 3.1). Characterizing small sets is facilitated by the topological T-chain property as formalized in the next section.

3 Main Results I: Irreducibility, aperiodicity, and T-chain property of normalized Markov chains underlying the CMA-ES algorithm

We present in this section one of the two main results of this paper stating the irreducibility, aperiodicity and T-chain property of the normalized chains underlying the CMA-ES algorithm defined in (2.14). We start by introducing the definitions of irreducibility, aperiodicity and T-kernel. Let P be a transition kernel on a state space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. We say that P is irreducible when there exists a nontrivial nonnegative measure φ on $\mathcal{B}(\mathsf{X})$ such that, for every $x \in \mathsf{X}$ and every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\varphi(\mathsf{A}) > 0$, there exists a positive integer k satisfying $P^k(x, \mathsf{A}) > 0$. When a measure φ satisfies this definition, we say that P is φ-irreducible.

When P is irreducible, the period of P is the largest integer $k \geqslant 1$ such that there exist disjoint sets $\mathsf{D}_1, \ldots, \mathsf{D}_k \in \mathcal{B}(\mathsf{X})$ with

$$\begin{cases} \varphi((\mathsf{D}_1 \cup \cdots \cup \mathsf{D}_k)^c) = 0 \text{ for every irreducibility measure } \varphi \text{ of } P \\ P(x_i, \mathsf{D}_{i+1}) = 1 \text{ for } x_i \in \mathsf{D}_i \text{ and } i = 0, \ldots, k-1 \ (\mathrm{mod}\ k). \end{cases} \tag{3.1}$$

An irreducible transition kernel P always admits a period $k \geqslant 1$ [35, Theorem 5.4.4], and when $k = 1$, P is said to be aperiodic.

For any positive integer m, a set $\mathsf{C} \in \mathcal{B}(\mathsf{X})$ is called m-small when there exists a nontrivial measure ν_m on $\mathcal{B}(\mathsf{X})$ such that $P^m(x, \mathsf{A}) \geqslant \nu_m(\mathsf{A})$ for every $x \in \mathsf{C}$ and every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$.

Given a probability distribution b on $\mathbb{N}$, we define the transition kernel K_b on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ as $K_b(x, \mathsf{A}) = \sum_{k\geqslant 0} b(k) P^k(x, \mathsf{A})$.

A substochastic kernel on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is a function $T \colon \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to \mathbb{R}$ such that $T(\cdot, \mathsf{A})$ is measurable for every $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ and $T(x, \cdot)$ is a finite measure on $\mathcal{B}(\mathsf{X})$ with $T(x, \mathsf{X}) \leqslant 1$ for every $x \in \mathsf{X}$. We say that the substochastic kernel T is a *continuous component* of the transition kernel K_b when $T(\cdot, A)$ is lower semicontinuous on X, $T(x, \mathsf{X}) > 0$ and $K_b(x, \mathsf{A}) \geqslant T(x, \mathsf{A})$ for every $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. A transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is called a T-kernel

when there exist a probability measure b on $\mathbb{N}$ and a substochastic kernel T which is a continuous component of the transition kernel K_b. Moreover, we say that a Markov chain is irreducible, respectively aperiodic, a T-chain, when its transition kernel is irreducible, respectively aperiodic, a T-kernel. We can now state our first main contribution presented in the next theorem and its corollary. They constitute a first milestone towards a linear convergence proof of CMA-ES. The complete proof of the following theorem is presented in Section 5 (cf. Theorem 5.1).

Theorem 3.1. *Suppose that the objective function f, the normalization function R, the stepsize change Γ and the sampling distribution ν_U^d satisfy* **F**1-**F**2, **R**1-**R**3, **Γ**1-**Γ**3 *and* **N**1, *respectively.*

Let $\Phi = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ be the normalized Markov chain underlying CMA-ES defined via (2.14) and P its transition kernel. Assume that $0 < c_1 + c_\mu < 1$. Then,

(i) if $c_c, c_\sigma \in (0,1)$, $c_\mu > 0$ and $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, then P is an irreducible aperiodic T-kernel, such that compact sets of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(ii) if $c_c \in (0,1)$, $c_\sigma = 1$ and $c_\mu > 0$, then the process $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ is an irreducible aperiodic T-chain, such that compact sets of $\mathbb{R}^d \times \mathbb{R}^d \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ are small;

(iii) if $c_\sigma \in (0,1)$ and $c_c = 1$, then the process $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ is an irreducible aperiodic T-chain, such that compact sets of $\mathbb{R}^d \times \mathbb{R}^d \times R^{-1}(\{1\})$ are small;

(iv) if $c_c = c_\sigma = 1$, then the process $\{(z_t, \mathbf{\Sigma}_t)\}_{t\geqslant 1}$ is an irreducible aperiodic T-chain, such that compact sets of $\mathbb{R}^d \times R^{-1}(\{1\})$ are small.

This result covers the entire range of eligible hyperparameter settings for CMA-ES except when $c_1 + c_\mu = 1$, or $c_\mu = 0$ and $c_c < 1$, or $1 - c_c = (1-c_\sigma)\sqrt{1-c_1-c_\mu} > 0$. Most importantly, when cumulation is used in the rank-one update, we need for our proof the rank-mu update. Without cumulation however ($c_c = 1$), the rank-one update is already sufficient to prove irreducibility and aperiodicity.

We finally formulate a particular case of Theorem 3.1. Using Proposition 2.1, Lemma 2.1 and Theorem 3.1, we find that Markov chains obtained with some standard stepsize change of CMA-ES and normalized by its minimum eigenvalue (possibly expressed in a different coordinate system which would be more fitted to the objective function f) or $\det(\cdot)^{1/d}$ are irreducible, aperiodic T-chains.

Corollary 3.1. *Let $\mathbf{H} \in \mathcal{S}_{++}^d$. Consider the process Φ defined via (2.14) with a normalization function $R = \det(\cdot)^{1/d}$ or $R = \lambda_{\min}(\mathbf{H}^{1/2} \times \cdot \times \mathbf{H}^{1/2})/\lambda_{\min}(\mathbf{H})$ and with the CSA stepsize change $\Gamma = \Gamma_{CSA}^1$ or $\Gamma = \Gamma_{CSA}^2$, see (2.9) or (2.10), respectively. Assume as in Theorem 3.1 that f satisfies* **F**1-**F**2 *and the sampling distribution satisfies* **N**1, *then under the same conditions on the hyperparameters as in Theorem 3.1, Φ is an irreducible aperiodic T-chain.*

4 Main results II: Extension of the analysis of nonlinear state-space models

We present in this section our methodological extensions of tools to analyze the irreducibility, aperiodicity and T-chain property of Markov chains. After reminding the basics in Section 4.1, we present two extensions.

First, some of the learning rate settings from Theorem 3.1(i), (ii), (iii) and (iv) give rise to a so-called *redundant Markov chain*, where one state variable can be dropped to define another Markov chain. We thus introduce redundant and projected Markov chains in Section 4.2 and explain how irreducibility, aperiodicity and T-chain property of the projected chain can be deduced from an analysis of the redundant chain. The main result of this section is Theorem 4.2.

The second methodological extension is motivated by the Markov chain (2.15) which is valued in a possibly nonsmooth manifold since the normalization $R(\cdot)$ may be not continuously differentiable, for instance when $R(\cdot) = \lambda_{\min}(\cdot)$. To analyze such a chain, we apply a homeomorphic transformation, thereby defining a Markov chain valued in a smooth manifold, and explain how irreducibility, aperiodicity and the T-chain property of the original Markov chain can be deduced from an analysis of the transformed Markov chain.

These results are applied in Section 5 for the proof of Theorem 3.1.

4.1 Deterministic control model and sufficient conditions for irreducibility and aperiodicity

We introduce in this section different definitions and theorems our analysis is based on the original article [19] to which we refer for more details. Let X and V be two smooth connected manifolds,[4] equipped with their Borel

[4] In the rest of the paper, manifolds will be considered as connected.

σ-fields, denoted $\mathcal{B}(\mathsf{X})$ and $\mathcal{B}(\mathsf{V})$, respectively. We later denote the dimension of X by n. The tangent space of X at a point $x \in \mathsf{X}$ is denoted $\mathrm{T}_x\mathsf{X}$, and we denote $\mathrm{dist}_{\mathsf{X}}$ and $\mathrm{dist}_{\mathsf{V}}$ the distance functions on X and V, respectively, which induce their respective topology. Consider a transition kernel P on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ associated to the Markov chain following the update equation

$$\phi_{t+1} = F(\phi_t, \alpha(\phi_t, U_{t+1})) \tag{4.1}$$

where $F\colon \mathsf{X} \times \mathsf{V} \to \mathsf{X}$ and $\alpha\colon \mathsf{X} \times \mathsf{U} \to \mathsf{V}$ are measurable functions, and $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an i.i.d. process independent of ϕ_0 and valued in a measurable space $(\mathsf{U}, \mathcal{U})$, where $\mathcal{U}$ is a σ-field of U.[5] We consider additionally the following assumptions on the model.

H1. *For any $x \in \mathsf{X}$, the distribution μ_x of the random variable $\alpha(x, U_1)$ admits a density p_x with respect to a σ-finite measure ζ_{V} on V, such that*

(i) the function $(x, v) \mapsto p_x(v)$ is lower semicontinuous;

(ii) for $\mathsf{A} \in \mathcal{B}(\mathsf{V})$, $\zeta_{\mathsf{V}}(\mathsf{A}) = 0$ if and only if A is negligible, i.e., $\mathrm{Leb}(\varphi(\mathsf{A} \cap V)) = 0$ for every local chart (φ, V) of V.

H2. *The function $F : \mathsf{X} \times \mathsf{V} \to \mathsf{X}$ is locally Lipschitz (with respect to the metrics $\mathrm{dist}_{\mathsf{X}} \oplus \mathrm{dist}_{\mathsf{V}}$ and $\mathrm{dist}_{\mathsf{X}}$).*

Below, Proposition 5.3 provides the Markov chain (5.3) that follows the control model (4.1) and satisfies **H**1 and **H**2 under mild assumptions on the objective function f and the stepsize change Γ.

We define inductively the *extended transition map* $S_x^k\colon \mathsf{V}^k \to \mathsf{X}$ associated to (4.1) for any $k \in \mathbb{N}$, $x \in \mathsf{X}$ and $v_{1:k} = (v_1, \dots, v_k) \in \mathsf{V}^k$ as follows

$$\begin{cases} S_x^0 \coloneqq x \\ S_x^k(v_{1:k}) \coloneqq F\left(S_x^{k-1}(v_{1:k-1}), v_k\right) & \text{for } k \geqslant 1. \end{cases} \tag{4.2}$$

From this definition, we obtain that if F is locally Lipschitz (respectively differentiable), then $(x, v_{1:k}) \mapsto S_x^k(v_{1:k})$ is locally Lipschitz (respectively differentiable). Likewise, we define the *extended probability density* $p_x^k\colon \mathsf{V}^k \to \mathbb{R}_+$ by

$$\begin{cases} p_x^1(v_1) \coloneqq p_x(v_1) \\ p_x^k(v_{1:k}) \coloneqq p_x^{k-1}(v_{1:k-1}) \times p_{S_x^{k-1}(v_{1:k-1})}(v_k) & \text{for } k \geqslant 2. \end{cases} \tag{4.3}$$

Given the Markov chain $\{\phi_t\}_{t\in\mathbb{N}}$ defined via (4.1), the function p_x^k is a density associated to the random variable $(\alpha(\phi_0, U_1), \dots, \alpha(\phi_{k-1}, U_k))$, when $\phi_0 = x$. For $x \in \mathsf{X}$ and $k \in \mathbb{N}^*$, we define the *control sets* of (4.1) by

$$\mathcal{O}_x^k \coloneqq \left\{v_{1:k} \in \mathsf{V}^k \mid p_x^k(v_{1:k}) > 0\right\} \ . \tag{4.4}$$

Assumption **H**1(i) implies that these sets are open subsets of V^k. We define moreover

$$\mathcal{O}_x^\infty \coloneqq \left\{v_{1:\infty} \in \mathsf{V}^{\mathbb{N}} \mid \forall k \geqslant 1, v_{1:k} \in \mathcal{O}_x^k\right\} \ . \tag{4.5}$$

We say that $x^* \in \mathsf{X}$ is a *steadily attracting state*, when for every $x \in \mathsf{X}$ and every neighborhood U of x^*, there exists $T > 0$ such that for every $k \geqslant T$, there exists $v_{1:k} \in \mathcal{O}_x^k$ such that $S_x^k(v_{1:k}) \in U$. When F is continuous, $v_{1:k}$ can be taken in $\overline{\mathcal{O}_x^k}$ [19, Corollary 4.5], i.e., $x^* \in \mathsf{X}$ is steadily attracting if and only if for every neighborhood U of x^*, there exists $T > 0$ such that for every $k \geqslant T$, there exists $v_{1:k} \in \overline{\mathcal{O}_x^k}$ such that $S_x^k(v_{1:k}) \in U$. In particular, if for every $x \in \mathsf{X}$, there exists $v_{1:\infty} \in \overline{\mathcal{O}_x^\infty}$ such that $S_x^k(v_{1:k})$ tends to x^*, then x^* is a steadily attracting state [19, Corollary 4.5].

We formulate now the following controllability condition of a steadily attracting state.

H3. *There exist a steadily attracting state $x^* \in \mathsf{X}$, an integer $k > 0$, and a path $v_{1:k}^* \in \overline{\mathcal{O}_{x^*}^k}$, such that $\partial S_{x^*}^k(v_{1:k}^*)$ is of maximal rank.*

For a locally Lipschitz function $G\colon \mathsf{V}^k \to \mathsf{X}$, $\partial G(v)$ is the Clarke's derivative of G at a point $v \in \mathsf{V}^k$, which is a set of linear applications between $\mathrm{T}_v\mathsf{V}^k$ and $\mathrm{T}_{G(v)}\mathsf{X}$ [19, Appendix B]. If G is differentiable at v, then $\partial G(v) = \{\mathcal{D}G(v)\}$, where $\mathcal{D}G(v)$ denotes the usual differential application of G in v. We then say that $\partial G(v)$ is of maximal rank when its elements are of maximal rank, that is, of rank n (the dimension of X). When G is differentiable at v and $\mathcal{D}G(v)$ is of maximal rank, then $\partial G(v) = \{\mathcal{D}G(v)\}$ is of maximal rank. We base our analysis on the following statement.

[5] Since we do not assume U to be a topological space, we consider a general σ-field $\mathcal{U}$ instead of its Borel σ-field.

Theorem 4.1 (Sufficient conditions for irreducibility and aperiodicity [19, Theorem 2.3])**.** *Consider the Markov kernel P defined via (4.1) such that* **H**1-**H**3 *are satisfied. Then P is an irreducible, aperiodic T-kernel, and every compact set of X is small.*

This theorem summarises the methodology we follow to analyze a normalized Markov chain underlying CMA-ES: we prove that the chain satisfies (4.1) as well as conditions **H**1-**H**3[6] under appropriate conditions on the learning rates, as well as on the functions f, Γ and R, and on the sampling distribution ν_U^d.

4.2 Irreducibility and aperiodicity of a projected Markov chain

The CMA-ES algorithm maintains two paths p_t^c and p_t^σ that do not parametrize the probability distribution for sampling candidate solutions but are used for (accelerating) the update of the covariance matrix and the stepsize, respectively. Yet, when no cumulation for the stepsize path is used, i.e., $c_\sigma = 1$, or no cumulation for the rank-one update path is used, i.e., $c_c = 1$, the CMA-ES algorithm typically still works properly while it is sometimes slower [18]. In these cases, the normalized Markov chain underlying CMA-ES can be described with fewer variables: p_t^c and p_t^σ boil down to random vectors that depend on the previous step only through the ranking permutation of candidate solutions. In order to analyze those algorithm variants without repeating proofs with small variations, we introduce here a method that allows to derive properties for a projected Markov chain from a redundant Markov chain with a specific parameter setting. We define a *redundant* Markov chain as a Markov chain $\{(\phi_t, \xi_t)\}_{t\in\mathbb{N}}$ valued in a topological product space $\mathsf{X}\times\mathsf{Y}$ such that the process $\{\phi_t\}_{t\in\mathbb{N}}$ also is a Markov chain, valued in X. In that case, we say that $\{\phi_t\}_{t\in\mathbb{N}}$ is a *projected* Markov chain of $\{(\phi_t, \xi_t)\}_{t\in\mathbb{N}}$.

Prior to that, we formalize the simplification of the normalized Markov chain of CMA-ES when at least one cumulation parameter is set to 1. The proof is a direct consequence of Proposition 2.2 and thus omitted.

Corollary 4.1. *Suppose that the objective function f and that the normalization function R satisfy* **F**2 *and* **R**1*, respectively. Let $\{(m_t, p_t^\sigma, p_t^c, \mathbf{C}_t, \sigma_t)\}_{t\in\mathbb{N}}$ be the Markov chain associated to CMA-ES defined in Section 2.1 and $\Phi = \{\phi_t\}_{t\geqslant 1} = \{(z_t, p_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ be the normalized process defined in (2.14).*

(i) If $c_\sigma = 1$, then the process $\{(z_t, q_t, \mathbf{\Sigma}_t, r_t)\}_{t\geqslant 1}$ defines a (time-homogeneous) Markov chain.

(ii) If $c_c = 1$, then the process $\{(z_t, p_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ defines a (time-homogeneous) Markov chain.

(iii) If $c_\sigma = c_c = 1$, then the process $\{(z_t, \mathbf{\Sigma}_t)\}_{t\in\mathbb{N}}$ defines a (time-homogeneous) Markov chain.

Corollary 4.1 motivates the introduction of the notion of a *projected chain* of a Markov chain $\tilde{\Phi}$, and to provide conditions for irreducibility and aperiodicity as in Theorem 4.1.

Define $\tilde{\Phi} = \{(\phi_t, \chi_t)\}_{t\in\mathbb{N}}$ a so-called *redundant* Markov chain on $(\mathsf{X}\times\mathsf{Y}, \mathcal{B}(\mathsf{X}\times\mathsf{Y}))$, with transition kernel $\tilde{P}$, such that

$$(\phi_{t+1}, \chi_{t+1}) = \tilde{F}(\phi_t, \chi_t, \tilde{\alpha}(\phi_t, \chi_t, U_{t+1})) \tag{4.6}$$

where $\tilde{F}\colon \mathsf{X}\times\mathsf{Y}\times\mathsf{V}\to\mathsf{X}\times\mathsf{Y}$ and $\tilde{\alpha}\colon \mathsf{X}\times\mathsf{Y}\times\mathsf{U}\to\mathsf{V}$ are measurable maps, $\mathsf{X}, \mathsf{Y}, \mathsf{V}$ are (smooth, connected) manifolds, $(\mathsf{U}, \mathcal{U})$ is a measurable space and $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an i.i.d. process valued in U, independent of (ϕ_0, χ_0). Assume **H**1-**H**2, and denote $\tilde{S}_{(x,y)}^k$, $\tilde{p}_{(x,y)}^k$ and $\tilde{\mathcal{O}}_{(x,y)}^k$ the extended transition map, the extended probability density and the control sets associated to the control model (4.6), for every $(x, y)\in\mathsf{X}\times\mathsf{Y}$ and $k\in\mathbb{N}$, respectively. Besides, we suppose redundancy of the chain by assuming that the function $\tilde{\alpha}$ does not depend on the variable χ, i.e., there exists a function α such that

$$\tilde{\alpha}(\phi, \chi, u) = \alpha(\phi, u) \quad \text{for every } \phi\in\mathsf{X}, \chi\in\mathsf{Y}, u\in\mathsf{U}. \tag{4.7}$$

Furthermore, we suppose that $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ is a Markov chain on X with transition kernel denoted P, following the next deterministic control model

$$\phi_{t+1} = F(\phi_t, \alpha(\phi_t, U_{t+1})), \tag{4.8}$$

with $\{U_{t+1}\}_{t\in\mathbb{N}}$ being the i.i.d. process introduced to define the redundant chain via (4.6). Then, we say that Φ is a projected chain of $\tilde{\Phi}$. As above, we denote S_x^k, p_x^k and $\mathcal{O}_x^k$ the extended transition map, the extended probability density and the control sets associated to the control model (4.8), for every $x\in\mathsf{X}$ and $k\in\mathbb{N}$, respectively. The next proposition connects the assumptions required for the two deterministic control models (4.6) and (4.8) that are useful to show that $\tilde{\Phi}$ and Φ are irreducible aperiodic T-chains.

[6] Condition **H**4 introduced in Section 4.2 is required instead of **H**3 when $c_c = 1$ or $c_\sigma = 1$.

Proposition 4.1. *Consider the control models associated to the redundant chain (4.6) and its associated projected chain (4.8). Then,*

(i) if **H1** *(resp.* **H2***) is satisfied for the redundant chain (4.6), then it is satisfied for its projected chain (4.8);*

(ii) the closures of the control sets $\mathcal{O}^k_\phi$ of the projected chain (4.8) equal the closures of the control sets $\tilde{\mathcal{O}}^k_{(\phi,\chi)}$ of the redundant chain (4.6), that is,

$$\overline{\mathcal{O}^k_\phi} = \overline{\tilde{\mathcal{O}}^k_{(\phi,\chi)}} \quad \text{for every } \phi \in \mathsf{X}, \chi \in \mathsf{Y}, k \geqslant 1; \tag{4.9}$$

(iii) the extended transition maps S^k_ϕ and $\tilde{S}^k_{(\phi,\chi)}$, defined in (4.2), of the control models of the projected chain (4.8), and the redundant chain (4.6), respectively, satisfy

$$S^k_\phi = \Pi_{\mathsf{X}} \circ \tilde{S}^k_{(\phi,\chi)} \quad \text{for every } \phi \in \mathsf{X}, \chi \in \mathsf{Y}, k \geqslant 1, \tag{4.10}$$

where $\Pi_{\mathsf{X}} \colon \mathsf{X} \times \mathsf{Y} \to \mathsf{X}$ is the canonical projection of $\mathsf{X} \times \mathsf{Y}$ on X.

Proof. First, we prove (i). Suppose that the redundant Markov chain following (4.6) satisfies **H1**, i.e., for all $(\phi, \chi) \in \mathsf{X} \times \mathsf{Y}$, the random variable $\tilde{\alpha}(\phi, \chi, U_1)$ admits a density $\tilde{p}_{(\phi,\chi)}$ with respect to a σ-finite measure ζ_{V} satisfying **H1**, such that $(\phi, \chi, v) \mapsto \tilde{p}_{(\phi,\chi)}(v)$ is lower semicontinuous. Let $\phi \in \mathsf{X}$. By (4.7) we have $\alpha(\phi, U_1) = \tilde{\alpha}(\phi, \chi, U_1)$ for every $\chi \in \mathsf{Y}$. Let $\chi_0 \in \mathsf{Y}$. Then the random variable $\tilde{\alpha}(\phi, \chi_0, U_1)$ admits a density $\tilde{p}_{(\phi,\chi_0)}$ with respect to a measure ζ_{V} on V satisfying **H1**(ii), and $(\phi, v) \mapsto \tilde{p}_{(\phi,\chi_0)}(v)$ is lower semicontinuous. Hence, $\alpha(\phi, U_1)$ also admits a density with respect to ζ_{V} denoted p_ϕ, such that $p_\phi(v) = \tilde{p}_{(\phi,\chi_0)}(v)$ for every $v \in \mathsf{V}$. By uniqueness of the density up to a null set, we obtain that, for $\phi \in \mathsf{X}$ and $\chi \in \mathsf{Y}$

$$p_\phi(v) = \tilde{p}_{(\phi,\chi)}(v) \quad \text{for } \zeta_{\mathsf{V}}\text{-almost every } v \in \mathsf{V}. \tag{4.11}$$

Since $(\phi, v) \mapsto p_{(\phi,\chi_0)}(v)$ is lower semicontinuous, we obtain that $(\phi, v) \mapsto p_\phi(v) = p_{(\phi,\chi_0)}(v)$ is lower semicontinuousand thus the projected chain satisfies **H1**. For assumption **H2**, if $\tilde{F}$ is locally Lipschitz, then, since by definition we have $F(\phi, v) = \Pi_{\mathsf{X}} \circ \tilde{F}(\phi, \chi, v)$ for $\phi \in \mathsf{X}$, $\chi \in \mathsf{Y}$ and $v \in \mathsf{V}$, by composition F is locally Lipschitz (Y being nonempty).

For (ii), from (4.11) and by definition of the control sets in (4.4), for every $(\phi, \chi) \in \mathsf{X} \times \mathsf{Y}$, $\mathcal{O}^k_\phi$ and $\tilde{\mathcal{O}}^k_{(\phi,\chi)}$ only differ by a ζ_{V}-negligible set. Besides, both are open sets and ζ_{V} is ,by **H1**(ii), a Borel measure, so $\overline{\mathcal{O}^k_\phi} = \overline{\tilde{\mathcal{O}}^k_{(\phi,\chi)}}$ (as a direct consequence of Carathéodory's criterion of Borel measures [15, Theorem 1.9]).

In order to prove (iii), we proceed by induction. Indeed, $S^0_\phi = \phi = \Pi_{\mathsf{X}}(\phi, \chi) = \Pi_{\mathsf{X}} \circ \tilde{S}^0_{(\phi,\chi)}$. Let $k \geqslant 0$ and assume $S^k_\phi = \Pi_{\mathsf{X}} \circ \tilde{S}^k_{(\phi,\chi)}$. Let $v_{1:k+1} \in \mathsf{V}^{k+1}$, we find $S^{k+1}_\phi(v_{1:k+1}) = F(S^k_\phi(v_{1:k}), v_{k+1}) = \Pi_{\mathsf{X}} \circ \tilde{F}(S^k_\phi(v_{1:k}), \chi_k, v_{k+1})$ where $\chi_k = \Pi_{\mathsf{Y}} \circ \tilde{S}^k_{(\phi,\chi)}(v_{1:k}) \in \mathsf{Y}$. By induction hypothesis we find that $\Pi_{\mathsf{X}} \circ \tilde{F}(S^k_\phi(v_{1:k}), \chi_k, v_{k+1}) = \Pi_{\mathsf{X}} \circ \tilde{F}(\tilde{S}^k_{(\phi,\chi)}(v_{1:k}), v_{k+1}) = \Pi_{\mathsf{X}} \circ \tilde{S}^{k+1}_{(\phi,\chi)}(v_{1:k+1})$ and thus $S^{k+1}_\phi(v_{1:k+1}) = \Pi_{\mathsf{X}} \circ \tilde{S}^{k+1}_{(\phi,\chi)}(v_{1:k+1})$. Hence $S^{k+1}_\phi = \Pi_{\mathsf{X}} \circ \tilde{S}^{k+1}_{(\phi,\chi)}$. □

We deduce the following result, which characterizes the controllability condition for the projected Markov chain.

Proposition 4.2. *Assume that F is continuous.*[7] *If $x^* = (\phi^*, \chi^*) \in \mathsf{X} \times \mathsf{Y}$ is a steadily attracting state of the redundant chain (4.6), then ϕ^* is a steadily attracting state of the projected chain (4.8).*

Suppose moreover that there exists $k \geqslant 1$ and $v^*_{1:k} \in \overline{\tilde{\mathcal{O}}^k_{x^*}}$ such that $\tilde{S}^k_{x^*}$ is differentiable at $v^*_{1:k}$, and for every $h^\phi \in \mathrm{T}_{\phi_k}\mathsf{X}$, there exists $h^\chi \in \mathrm{T}_{\chi_k}\mathsf{Y}$, where $(\phi_k, \chi_k) = \tilde{S}^k_{x^*}(v^*_{1:k})$ and with $(h^\phi, h^\chi) \in \operatorname{rge} \mathcal{D}\tilde{S}^k_{x^*}(v^*_{1:k})$. Then $v^*_{1:k} \in \overline{\mathcal{O}^k_{\phi^*}}$ and $\mathcal{D}S^k_{\phi^*}(v^*_{1:k})$ exists and is of maximal rank.

Proof. Since F is continuous, we can use the definition of a steadily attracting set via taking the elements for the k-step paths within the closure of the control sets (see Section 4.1) instead of the control sets. According to the previous proposition $\overline{\mathcal{O}^k_\phi} = \overline{\tilde{\mathcal{O}}^k_{(\phi,\chi)}}$ for every $\phi \in \mathsf{X}, \chi \in \mathsf{Y}, k \geqslant 1$ and we can thus easily prove that ϕ^* is steadily attracting for (4.8) when (ϕ^*, χ^*) is steadily attracting for (4.6). Moreover, by Proposition 4.1(iii), we have that $S^k_{\phi^*} = \Pi_{\mathsf{X}} \circ \tilde{S}^k_{(\phi^*,\chi^*)}$. Then, by the chain rule [13, Corollary 2.6.6], we have

$$\begin{aligned}\mathcal{D}S^k_{\phi^*}(v^*_{1:k}) &= \mathcal{D}\Pi_{\mathsf{X}}(\tilde{S}^k_{(\phi^*,\chi^*)}(v^*_{1:k})) \circ \mathcal{D}\tilde{S}^k_{(\phi^*,\chi^*)}(v^*_{1:k}) \\ &= \Pi_{\mathrm{T}_{S^k_{\phi^*}(v^*_{1:k})}\mathsf{X}} \circ \mathcal{D}\tilde{S}^k_{(\phi^*,\chi^*)}(v^*_{1:k}).\end{aligned}$$

[7] Alternatively, we can assume without loss of generality that for every $\phi \in \mathsf{X}$, the density functions $\tilde{p}_{(\phi,\chi)}$ are identical for $\chi \in \mathsf{Y}$.

Therefore every $h^\phi \in \mathrm{T}_{S^k_{\phi^*}(v^*_{1:k})}\mathsf{X}$ belongs to the range of $\mathcal{D}S^k_{\phi^*}(v^*_{1:k})$ (we use here that by assumption there exists $h^\chi \in \mathrm{T}_{\Pi_\mathsf{Y} S_{(\phi^*,\xi^*)}(v^*_{1:k})}$), making it a surjective linear map, hence of maximal rank. □

As a consequence, we derive sufficient conditions for irreducibility and aperiodicity of the kernel P of the projected Markov chain Φ. We replace the controllability condition **H**3 by the following.

H4. *There exist a steadily attracting x^*, an integer $k > 0$ and a path $v^*_{1:k} \in \overline{\tilde{\mathcal{O}}^k_{x^*}}$ such that $S^k_{x^*}$ is differentiable at $v^*_{1:k}$, and for every $h^\phi \in \mathrm{T}_\phi\mathsf{X}$, there exists $h^\chi \in \mathrm{T}_\chi\mathsf{Y}$ with $(h^\phi, h^\chi) \in \operatorname{rge} \mathcal{D}\tilde{S}^k_{x^*}(v^*_{1:k})$, where $(\phi, \chi) = \tilde{S}^k_{x^*}(v^*_{1:k})$.*

Theorem 4.2 (Sufficient conditions for irreducibility and aperiodicity of a projected Markov chain)**.** *Consider the control model of the redundant chain (4.6) and assume it satisfies conditions* **H**1-**H**2 *and* **H**4*. Then the kernel P of the projected chain defined via (4.8) is an irreducible aperiodic T-kernel, and every compact set of X is small.*

Proof. Denote $x^* = (\phi^*, \chi^*)$. Since x^* is steadily attracting for (4.6) and since F is continuous by **H**2, then by Proposition 4.2 ϕ^* is steadily attracting for (4.8). Besides, by Proposition 4.1(ii), we have $v^*_{1:k} \in \overline{\mathcal{O}^k_{x^*}} = \overline{\tilde{\mathcal{O}}^k_{\phi^*}}$. By Proposition 4.2 again, we find that $\mathcal{D}S^k_{\phi^*}(v^*_{1:k})$ exists and is of maximal rank. We complete the proof by applying Theorem 4.1. □

Theorem 4.2 is used later to analyze the normalized chain defined in (2.14) when $c_c = 0$ or $c_\sigma = 0$. Even though Theorem 4.1 could be applied directly to the projected chain, this generalization allows to find a steadily attracting state and prove a controllability condition on the same chain for all settings without repeating the same proof.

4.3 Homeomorphic transformation of an irreducible aperiodic T-chain

The state space of the chain Φ defined via (2.14) is not a smooth manifold if the normalization function R is not continuously differentiable on $\mathcal{S}^d_{++}$. In order to include nonsmooth functions R in our analysis (for instance if $R(\cdot)$ is the minimal eigenvalue of a positive definite matrix), we apply Theorem 4.1 (or Theorem 4.2 when we do not have cumulation) to a Markov chain Θ, defined as a homeomorphic transformation of Φ, such that the state space of Θ is a smooth manifold. This is achieved in Sections 5.1 to 5.4. Now, we explain why it is sufficient to prove that the transformed chain Θ is an irreducible, aperiodic T-chain to have the same properties on Φ.

Theorem 4.3. *Let $\xi\colon \mathsf{Y} \to \mathsf{X}$ be a homeomorphism between the topological spaces Y and X, equipped with their respective Borel σ-fields. Let $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ be a (time-homogeneous) Markov chain with state space Y, and define $\Theta = \{\xi(\phi_t)\}_{t\in\mathbb{N}}$. Then,*

(i) Θ is a (time-homogeneous) Markov chain with state space X;

(ii) if Θ is irreducible (resp. aperiodic, a T-chain), then Φ is irreducible (resp. aperiodic, a T-chain).

Proof. First, we prove (i). Denote P the Markov kernel of Φ. If the distribution of $\xi(\phi_0)$ is δ_x for $x \in \mathsf{X}$, then ϕ_0 is distributed under $\delta_{\xi^{-1}(x)}$. For $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ and $x \in \mathsf{X}$, we have then

$$\mathbb{P}\left[\xi(\phi_1) \in \mathsf{A} \mid \xi(\phi_0) = x\right] = \mathbb{P}\left[\phi_1 \in \xi^{-1}(\mathsf{A}) \mid \phi_0 = \xi^{-1}(x)\right] = P(\xi^{-1}(x), \xi^{-1}(\mathsf{A}))$$

Moreover, $P(\xi^{-1}(\cdot), \xi^{-1}(\cdot))$ defines a Markov kernel for Θ. Indeed, since ξ is a homeomorphism, ξ^{-1} is continuous and thus measurable. In particular the k-step transition kernel of Θ equals $P^k(\xi^{-1}(\cdot), \xi^{-1}(\cdot))$ (where P^k is the k-step transition kernel of Φ). Thus Θ is a time-homogeneous Markov chain.

Now we prove (ii). Suppose that Θ is irreducible, i.e., the kernel $P(\xi^{-1}(\cdot), \xi^{-1}(\cdot))$ admits a nontrivial nonnegative measure ϑ on $\mathcal{B}(\mathsf{X})$ such that for $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{B}(\mathsf{X})$ with $\vartheta(\mathsf{A}) > 0$, there exists $k > 0$ with $P^k(\xi^{-1}(x), \xi^{-1}(\mathsf{A})) > 0$. Then, for every $\mathsf{B} \in \mathcal{B}(\mathsf{Y})$ such that $\vartheta(\xi(\mathsf{B})) > 0$ and for every $y \in \mathsf{Y}$, there exists $k > 0$ with $P^k(x, \mathsf{B}) > 0$, i.e., Φ is $\vartheta \circ \xi$-irreducible. Likewise, for every irreducibility measure φ of Φ, then $\varphi \circ \xi^{-1}$ is a irreducibility measure of Θ. In particular, every irreducibility measure ϑ of Θ can be defined as $\varphi \circ \xi^{-1}$ for some irreducibility measure φ of Φ. Moreover, denote $k \geqslant 1$ the period of Θ, i.e., k is the largest integer such that there exists disjoint sets $\mathsf{D}_1, \dots, \mathsf{D}_k \in \mathcal{B}(\mathsf{Y})$ with

$$\begin{cases} \varphi((\mathsf{D}_1 \cup \cdots \cup \mathsf{D}_k)^c) = 0 \quad \text{for any irreducibility measure } \varphi \text{ of } \Phi \\ P(\xi^{-1}(y_i), \xi^{-1}(\mathsf{D}_{i+1})) = 1 \text{ for } y_i \in \mathsf{D}_i \text{ and } i = 0, \dots, k-1 \mod k. \end{cases}$$

Therefore k is the largest integer such that there exists disjoint sets $\mathsf{C}_1, \dots, \mathsf{C}_k \in \mathcal{B}(\mathsf{X})$ with

$$\begin{cases} \varphi(\xi^{-1}(\mathsf{C}_1 \cup \cdots \cup \mathsf{C}_k)^c) = 0 \quad \text{for any irreducibility measure } \varphi \text{ of } \Phi \\ P(x_i, \mathsf{C}_{i+1}) = 1 \text{ for } x_i \in \mathsf{C}_i \text{ and } i = 0, \dots, k-1 \mod k. \end{cases}$$

Hence, the period of Φ equals k the period of Θ. In particular, if Θ is aperiodic, then Φ is aperiodic.

Suppose now that Θ is a T-chain and let $T\colon \mathsf{X}\times\mathcal{B}(\mathsf{X}) \to \mathbb{R}_+$ be a substochastic kernel such that $K_b(\xi^{-1}(\cdot), \xi^{-1}(\cdot)) \geqslant T$ for some probability distribution b on $\mathbb{N}$, $T(\cdot, \mathsf{X}) > 0$ and $x \mapsto T(x, \mathsf{A})$ is lower semicontinuousfor $\mathsf{A} \in \mathcal{B}(\mathsf{X})$. Then, if we define $T'(y, \mathsf{B}) = T(\xi(y), \xi(\mathsf{B}))$ for $y \in \mathsf{Y}$ and $\mathsf{B} \in \mathcal{B}(\mathsf{Y})$, we obtain that T' is a substochastic kernel such that $K_b \geqslant T'$ for some probability distribution b on $\mathbb{N}$, $T'(\cdot, \mathsf{Y}) > 0$ and that $y \mapsto T(y, \mathsf{B})$ is lower semicontinuousfor every $\mathsf{B} \in \mathcal{B}(\mathsf{Y})$. Therefore, Φ is a T-chain. □

5 Proof of Theorem 3.1

The objective of this section is to prove Theorem 3.1. To do so, we investigate nonlinear state-space models associated to the recursion (2.15) using the theoretical tools presented in Section 4.

Since the normalization function R is not assumed to be smooth, we consider a transformed Markov chain—for which we can apply Theorem 4.3—valued in a smooth manifold and which can be transformed via a homeomorphism into the normalized Markov chain (2.15). In Section 5.1, we introduce the control model associated to this transformed process and verify the conditions **H**1, **H**2 reminded in Section 4.1.

Then, a last condition, **H**3 or **H**4,[8] is proven in two steps: Section 5.2 proves the existence of a steadily attracting state (defined in Section 4.1) and Section 5.3 shows that a required controllability condition is satisfied. We conclude the proof in Section 5.4, where the Markov chains associated to the different learning rate settings are analyzed, based on Theorem 4.2.

5.1 Definition of normalized chains underlying CMA-ES following (4.1) and satisfying H1-H2

In order to apply Theorem 4.1 to the normalized CMA-ES Markov chain defined via (2.14), we require the state space $\mathsf{Y} = \mathbb{R}^{3d} \times R^{-1}(\{1\}) \times \mathbb{R}_{++}$ to be a smooth connected manifold. As mentioned in the previous section, this is not necessarily true unless we assume that the normalization R is continuously differentiable. Hence, we introduce a homeomorphic transformation of the normalized chain which lives on a smooth manifold. Consider $\rho\colon \mathcal{S}^d_{++} \to \mathbb{R}_{++}$ a map satisfying

***ρ*1.** *the function ρ is (positively) homogeneous and $\rho(\mathbf{I}_d) = 1$,*

***ρ*2.** *the function ρ is smooth ($\mathcal{C}^\infty$) on $\mathcal{S}^d_{++}$.*

We keep this smooth normalization function abstract for the moment and will take it equal to $\rho(\cdot) = \det(\cdot)^{1/d}$ for proving Theorem 3.1. Define now

$$\begin{aligned} \xi\colon \quad & \mathsf{Y} \to \mathsf{X} \\ & (z, p, q, \mathbf{\Sigma}, r) \mapsto \left(z, p, q, \rho(\mathbf{\Sigma})^{-1}\mathbf{\Sigma}, r\right) \end{aligned} \tag{5.1}$$

where $\mathsf{X} = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$. Then, as stated in the next proposition, X defines a smooth connected manifold.

Proposition 5.1. *Suppose that the map $\rho\colon \mathcal{S}^d_{++} \to \mathbb{R}_{++}$ satisfies **ρ**1-**ρ**2. Then, the set $\mathsf{X} = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$ is a smooth connected manifold of dimension $3d + d(d+1)/2$.*

Proof. First note that the set $\mathsf{M} \coloneqq \mathbb{R}^{3d} \times \mathcal{S}^d_{++} \times \mathbb{R}_{++}$ is an open subset of the Euclidean space $\mathbb{R}^{3d} \times \mathcal{S}^d \times \mathbb{R}$, hence is a smooth submanifold of dimension $3d + d(d+1)/2 + 1$. Moreover, $\mathsf{N} \coloneqq \mathbb{R}$ is a smooth manifold of dimension 1. Define then the map

$$\begin{aligned} \bar{\varrho}\colon \quad & \mathsf{M} \to \mathsf{N} \\ & (z, p, q, \mathbf{\Sigma}, r) \mapsto \rho(\mathbf{\Sigma}). \end{aligned}$$

Then, by **ρ**2, $\bar{\varrho}$ is smooth. Moreover, it is a submersion at every point of M. Indeed, let $(z, p, q, \mathbf{\Sigma}, r) \in \mathsf{M}$, and let $\varepsilon \in (-1, 1)$. Then,

$$\bar{\varrho}\left((z, p, q, \mathbf{\Sigma}, r) + (0, 0, 0, \varepsilon\mathbf{\Sigma}, 0)\right) = \rho((1+\varepsilon)\mathbf{\Sigma}) = \rho(\mathbf{\Sigma}) + \varepsilon\rho(\mathbf{\Sigma})$$

by **ρ**1. Therefore, by Taylor expansion and since the derivative $\mathcal{D}\bar{\varrho}(z, p, q, \mathbf{\Sigma}, r)$ is linear, we have

$$\mathcal{D}\bar{\varrho}(z, p, q, \mathbf{\Sigma}, r)(0, 0, 0, \kappa\mathbf{\Sigma}, 0) = \kappa\rho(\mathbf{\Sigma}),$$

[8] **H**3 for case (i) and **H**4 for cases (ii), (iii), (iv) of Theorem 3.1.

for every $\kappa \in \mathbb{R}$, with $\rho(\boldsymbol{\Sigma}) > 0$. Hence, $\mathcal{D}\bar{\varrho}(z,p,q,\boldsymbol{\Sigma},r)\colon \mathbb{R}^{3d}\times\mathcal{S}^d\times\mathbb{R}\to\mathbb{R}$ is surjective and thus $\bar{\varrho}$ is a submersion. Therefore, by the submersion level set theorem [31, Corollary 5.14], $\mathsf{X} = \bar{\varrho}^{-1}(\{1\})$ is a smooth manifold of dimension $3d + d(d+1)/2 + 1 - 1 = 3d + d(d+1)/2$.

Let us prove now that X is connected. Since $\mathbb{R}^{3d}\times\mathbb{R}_{++}$ is connected, it is sufficient to prove that the manifold $\rho^{-1}(\{1\})$ is connected, and thus sufficient to prove that $\rho^{-1}(\{1\})$ is path-connected [31, Proposition 1.11]. Let $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \in \rho^{-1}(\{1\})$. Since $\mathcal{S}^d_{++}$ is connected, there exists a continuous path $\gamma\colon[0,1]\to\mathcal{S}^d_{++}$ with $\gamma(0) = \boldsymbol{\Sigma}_0$ and $\gamma(1) = \boldsymbol{\Sigma}_1$. Define then the path $\hat{\gamma}\colon[0,1]\to\rho^{-1}(\{1\})$ by $\hat{\gamma}(t) = \gamma(t)/\rho(\gamma(t))$ for $t\in[0,1]$. Since γ and ρ are continuous, then $\hat{\gamma}$ is continuous. Besides, $\hat{\gamma}(0) = \boldsymbol{\Sigma}_0/\rho(\boldsymbol{\Sigma}_0) = \boldsymbol{\Sigma}_0$ and $\hat{\gamma}(1) = \boldsymbol{\Sigma}_1/\rho(\boldsymbol{\Sigma}_1) = \boldsymbol{\Sigma}_1$, ending the proof. □

Moreover, the map ξ defined in (5.1) is a homeomorphism as stated below.

Proposition 5.2. *Suppose that R and ρ are both continuous and satisfy* **R**1 *and* **ρ**1*. Then, the map ξ defined in (5.1) between the sets Y and X is a homeomorphism and*

$$\begin{aligned}\xi^{-1}\colon\quad &\mathsf{X}\to\mathsf{Y}\\ &(z,p,q,\hat{\boldsymbol{\Sigma}},r)\mapsto\left(z,p,q,R(\hat{\boldsymbol{\Sigma}})^{-1}\hat{\boldsymbol{\Sigma}},r\right).\end{aligned}\tag{5.2}$$

Proof. We can easily verify that the expression of the reciprocal function of ξ is (5.2). Then ξ^{-1} (resp. ξ) is continuous since R (resp. ρ) is continuous and takes value in $\mathbb{R}_{++}$. □

We formalize in the next lemma the update equations for $\{\theta_t\}_{t\in\mathbb{N}} = \{\xi(\phi_t)\}_{t\in\mathbb{N}}$.

Lemma 5.1. *Suppose that the normalization function R satisfies* **R**1 *and let $\Phi = \{\phi_t\}_{t\in\mathbb{N}}$ be the Markov chain defined via (2.15). Let ρ be a normalization function satisfying* **ρ**1 *and let ξ be the homeomorphism defined in (5.1). Then the Markov chain $\Theta = \{\theta_t\}_{t\in\mathbb{N}} = \{\xi(\phi_t)\}_{t\in\mathbb{N}}$ satisfies*

$$\begin{aligned}
z_{t+1} &= \frac{F^m_{c_m}(z_t, \sqrt{R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}})}{\sqrt{r_{t+1}}\Gamma(p_{t+1})}\\
p_{t+1} &= F^p_{c_\sigma}(p_t, \mathbf{w}_m^\top U_{t+1}^{s_{t+1}})\\
q_{t+1} &= F^p_{c_c}(r_t^{-1/2}q_t, \sqrt{R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t}\mathbf{w}_m^\top U_{t+1}^{s_{t+1}})\\
\hat{\boldsymbol{\Sigma}}_{t+1} &= \frac{F^C_{c_1,c_\mu}\left(R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t, q_{t+1}, R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\sqrt{\hat{\boldsymbol{\Sigma}}_t}\sum_{i=1}^\mu w_i^c\left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top\sqrt{\hat{\boldsymbol{\Sigma}}_t}\right)}{\rho\circ F^C_{c_1,c_\mu}\left(R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t, q_{t+1}, R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\sqrt{\hat{\boldsymbol{\Sigma}}_t}\sum_{i=1}^\mu w_i^c\left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top\sqrt{\hat{\boldsymbol{\Sigma}}_t}\right)}\\
r_{t+1} &= R\circ F^C_{c_1,c_\mu}\left(R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t, q_{t+1}, R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\sqrt{\hat{\boldsymbol{\Sigma}}_t}\sum_{i=1}^\mu w_i^c\left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top\sqrt{\hat{\boldsymbol{\Sigma}}_t}\right)
\end{aligned}\tag{5.3}$$

where $\{U_{t+1}\}_{t\in\mathbb{N}}$ is an i.i.d. process independent of $\theta_0 = (z_0,p_0,q_0,\hat{\boldsymbol{\Sigma}}_0,r_0) = \xi(\phi_0)$ distributed in the measured space $\mathsf{U} = (\mathbb{R}^d)^\lambda$ with $U_1\sim\nu_U^d$, and s_{t+1} is a permutation of $\mathfrak{S}_\lambda$ that sorts the f-values of $x^* + z_t + \sqrt{R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t}U_{t+1}^i$, for $i = 1,\dots,\lambda$.

Proof. Since for $t\in\mathbb{N}$, according to (5.2), we have that $\boldsymbol{\Sigma}_t = R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t$, the update equations for p_{t+1}, q_{t+1} ,r_{t+1} and z_{t+1} in (5.3) are deduced directly from Proposition 2.2 where we replace $\boldsymbol{\Sigma}_t$ by $R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t$. Moreover, we have, by Proposition 2.2 and using the definition of $\tilde{\boldsymbol{\Sigma}}_{t+1}$ in (2.16)

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{t+1} = \frac{\boldsymbol{\Sigma}_{t+1}}{\rho(\boldsymbol{\Sigma}_{t+1})} &= \frac{\tilde{\boldsymbol{\Sigma}}_{t+1}}{R(\tilde{\boldsymbol{\Sigma}}_{t+1})}\times\frac{R(\tilde{\boldsymbol{\Sigma}}_{t+1})}{\rho(\tilde{\boldsymbol{\Sigma}}_{t+1})}\\
&= \frac{F^{\mathrm{C}}_{c_1,c_\mu}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^\mu w_i^c\left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top\sqrt{\boldsymbol{\Sigma}_t}\right)}{\rho\circ F^{\mathrm{C}}_{c_1,c_\mu}\left(\boldsymbol{\Sigma}_t, q_{t+1}, \sqrt{\boldsymbol{\Sigma}_t}\sum_{i=1}^\mu w_i^c\left[U_{t+1}^{s_{t+1}(i)}\right]\left[U_{t+1}^{s_{t+1}(i)}\right]^\top\sqrt{\boldsymbol{\Sigma}_t}\right)}.
\end{aligned}$$

The proof ends by replacing $\boldsymbol{\Sigma}_t$ by $R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t$. □

Using Theorem 4.3, we can transfer the irreducibility, aperiodicity and the T-chain property from the Markov chain Θ to the original normalized chain Φ we are interested in. Our objective from now on is to prove that the Markov chain Θ is an irreducible aperiodic T-chain. Our strategy for that is to apply Theorem 4.1 and verify that

the required assumptions are satisfied. We first prove that Θ follows a deterministic control model of the form (4.1), described in Section 4.1.

Consider the smooth manifold $\mathsf{X} = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$ (see Proposition 5.1) that defines the state space of the Markov chain Θ and let $\mathsf{V} \coloneqq \mathbb{R}^{d\mu}$. We define

$$\begin{aligned} \alpha_\Theta \colon \quad \mathsf{X} \times \mathsf{U} &\to \mathsf{V} \\ ((z,p,q,\hat{\mathbf{\Sigma}},r),(u^1,\dots,u^\lambda)) &\mapsto \left[\sqrt{\frac{\hat{\mathbf{\Sigma}}}{R(\hat{\mathbf{\Sigma}})}} u^{s^u_{f(x^*+z+\sqrt{R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}}}\cdot)}}\right]^{(i)}_{i=1,\dots,\mu} \end{aligned} \tag{5.4}$$

where given $g \colon \mathsf{A} \to \mathbb{R}$ a function and $v \in \mathsf{A}^\lambda$ we have used the notation s^v_g for a permutation that sorts increasingly the $g(v^i)$, $i = 1,\dots,\lambda$. Consider the $(z^+, p^+, q^+, \hat{\mathbf{\Sigma}}^+, r^+)$ the update of $\theta = (z,p,q,\hat{\mathbf{\Sigma}},r)$ given the random input equals $v = \alpha_\Theta(\theta,u)$ that is

$$z^+ = \frac{z + c_m \mathbf{w}_m^\top v}{\sqrt{r^+ \Gamma(p^+)}} \tag{5.5}$$

$$p^+ = (1-c_\sigma)p + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}} R(\hat{\mathbf{\Sigma}})^{1/2} \hat{\mathbf{\Sigma}}^{-1/2} \mathbf{w}_m^\top v \tag{5.6}$$

$$q^+ = r^{-1/2}(1-c_c)q + \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \mathbf{w}_m^\top v \tag{5.7}$$

$$\hat{\mathbf{\Sigma}}^+ = \frac{(1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}} + c_1 q^+ (q^+)^\top + c_\mu \sum_{i=1}^\mu w_i^c v_i v_i^\top}{\rho\left((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}} + c_1 q^+ (q^+)^\top + c_\mu \sum_{i=1}^\mu w_i^c v_i v_i^\top\right)} \tag{5.8}$$

$$r^+ = R\left((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}} + c_1 q^+ (q^+)^\top + c_\mu \sum_{i=1}^\mu w_i^c v_i v_i^\top\right). \tag{5.9}$$

This update defines a function $F_\Theta \colon \mathsf{X} \times \mathsf{V} \to \mathsf{X}$ such that

$$(z^+, p^+, q^+, \hat{\mathbf{\Sigma}}^+, r^+) = F_\Theta((z,p,q,\hat{\mathbf{\Sigma}},r), \alpha_\Theta(\theta,u))$$

that can be expressed as

$$\begin{aligned} F_\Theta((z,p,q,\hat{\mathbf{\Sigma}},r),(v^1,\dots,v^\mu)) = \Big(&F_z(z,p,q,R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}},r;v), F_p(p,R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}};v), F_q(q,r;v), \\ &F_{\mathbf{\Sigma}}(q,R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}},r;v), F_r(q,R(\hat{\mathbf{\Sigma}})^{-1}\hat{\mathbf{\Sigma}},r;v)\Big)^\top \end{aligned} \tag{5.10}$$

for $(z,p,q,\hat{\mathbf{\Sigma}},r) \in \mathsf{X}$ and $(v^1,\dots,v^\mu) \in \mathsf{V}$, and where F_z, F_p, F_q, $F_{\mathbf{\Sigma}}$ and F_r are defined as follows

$$F_z(z,p,q,\mathbf{\Sigma},r;v) = F_r(q,\mathbf{\Sigma},r;v)^{-1/2}\Gamma \circ F_p(p,\mathbf{\Sigma};v)^{-1} F^{\text{m}}_{c_m}(z, \mathbf{w}_m^\top v) \tag{5.11}$$

$$F_p(p,\mathbf{\Sigma};v) = F^{\text{p}}_{c_\sigma}(p, \mathbf{\Sigma}^{-1/2}\mathbf{w}_m^\top v) \tag{5.12}$$

$$F_q(q,r;v) = F^{\text{p}}_{c_c}(r^{-1/2}q, \mathbf{w}_m^\top v) \tag{5.13}$$

$$F_{\mathbf{\Sigma}}(q,\mathbf{\Sigma},r;v) = \frac{F^{\text{C}}_{c_1,c_\mu}\left(\mathbf{\Sigma}, F_q(q,r;v), \sum_{i=1}^\mu w_i^c v_i v_i^\top\right)}{\rho \circ F^{\text{C}}_{c_1,c_\mu}\left(\mathbf{\Sigma}, F_q(q,r;v), \sum_{i=1}^\mu w_i^c v_i v_i^\top\right)} \tag{5.14}$$

$$F_r(q,\mathbf{\Sigma},r;v) = R \circ F^{\text{C}}_{c_1,c_\mu}\left(\mathbf{\Sigma}, F_q(q,\mathbf{\Sigma},r;v), \sum_{i=1}^\mu w_i^c v_i v_i^\top\right). \tag{5.15}$$

Then, as stated in the next proposition, Θ follows the model described in Section 4.1 with the functions F_Θ and α_Θ defined above.

Proposition 5.3. *Suppose that the normalization functions R satisfies* **R**1 *and ρ satisfies* $\boldsymbol{\rho}$1*. Then, the Markov chain $\Theta = \{(z_t,p_t,q_t,\hat{\mathbf{\Sigma}}_t,r_t)\}_{t\in\mathbb{N}}$ defined by (5.3) satisfies*

$$\theta_{t+1} = F_\Theta\left(\theta_t, \alpha_\Theta(\theta_t, U_{t+1})\right) \tag{5.16}$$

where F_Θ is defined in (5.10) and α_Θ in (5.4).

Proof. Straightforward by Lemma 5.1. □

Before to prove that the control model (5.16) associated to the Markov chain Θ satisfies the assumptions **H**1 and **H**2, we characterize in the next lemma the density of the random variable $\alpha_\Theta(\theta, U)$ for $\theta \in \mathsf{X}$ and $U \sim (\nu_U^d)^{\otimes\lambda}$ assuming that the objective function is the composite of a strictly increasing function with a function with negligible level sets and the distribution ν_U^d is admits a density positive everywhere with respect to the Lebesgue measure. The latter assumption could be relaxed with more work, but for the purposes of this paper we only consider positive densities.

Lemma 5.2. *Suppose that the objective function f satisfies* **F**1 *and that the probability distribution ν_U^d satisfies* **N**1. *Define, for any $\theta = (z, p, q, \hat{\mathbf{\Sigma}}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{S}_{++}^d \times \mathbb{R}_{++}$ and $v = (v_1, \ldots, v_\mu) \in \mathbb{R}^{d\mu}$,*

$$p_{z,\mathbf{\Sigma}}(v) = \frac{\lambda!}{(\lambda-\mu)!} \mathbb{1}\{f_*(z+\sqrt{\mathbf{\Sigma}}v_1) < \cdots < f_*(z+\sqrt{\mathbf{\Sigma}}v_\mu)\}(1 - Q_{z,\mathbf{\Sigma}}^{f_*}(v_\mu))^{\lambda-\mu} p_U^d(v_1)\ldots p_U^d(v_\mu) \tag{5.17}$$

with $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}/R(\hat{\mathbf{\Sigma}})$ where R is the normalization function used in (5.4), $Q_{z,\mathbf{\Sigma}}^{f_*}(u) = \int \mathbb{1}\{f_*(z+\sqrt{\mathbf{\Sigma}}\xi) < f_*(z+\sqrt{\mathbf{\Sigma}}u)\}\nu_U^d(\mathrm{d}\xi)$ for $u \in \mathbb{R}^d$, and $f_* = f(\cdot + x^*)$. Then, $p_{z,\mathbf{\Sigma}}$ defines a density (with respect to Lebesgue in $\mathbb{R}^{d\mu}$) of the random variable $\mathbf{\Sigma}^{-1/2}\alpha_\Theta(\theta, U)$, where $U \sim (\nu_U^d)^{\otimes\lambda}$ such that the density of $\alpha_\Theta(\theta, U)$ equals

$$v \mapsto \det \mathbf{\Sigma}^{-1/2} p_{z,\mathbf{\Sigma}}(\mathbf{\Sigma}^{-1/2}v) = \frac{1}{\sqrt{\det \mathbf{\Sigma}}} p_{z,\mathbf{\Sigma}}(\mathbf{\Sigma}^{-1/2}v) \ . \tag{5.18}$$

Besides, when R is continuous, the function $((z, p, q, \hat{\mathbf{\Sigma}}, r), v) \in \mathbb{R}^{3d} \times \mathcal{S}_{++}^d \times \mathbb{R}_{++} \times \mathbb{R}^{d\mu} \mapsto p_{z,\hat{\mathbf{\Sigma}}/R(\hat{\mathbf{\Sigma}})}(v)$ is lower semicontinuous and thus the density function (5.18) is lower semicontinuous as well.

Proof. Let $U^1, \ldots, U^\lambda$ be independent random vectors identically distributed under the probability distribution ν_U^d, and denote $U = (U^1, \ldots, U^\lambda)$. Let $\theta = (z, p, q, \hat{\mathbf{\Sigma}}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{S}_{++}^d \times \mathbb{R}_{++}$. Since the objective function f satisfies **F**1, then the random vector $V = \mathbf{\Sigma}^{-1/2}\alpha_\Theta(\theta, U)$ satisfies almost surely

$$V = \sum_{\sigma \in \mathfrak{S}_\lambda} \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(1)}\right) < \cdots < f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(\lambda)}\right)\right\} \times \left(U^{\sigma(1)}, \ldots, U^{\sigma(\mu)}\right).$$

where $\mathfrak{S}_\lambda$ is the set of permutations of $\{1, \ldots, \lambda\}$. Hence, by symmetry,

$$V = \frac{1}{(\lambda-\mu)!} \sum_{\sigma \in \mathfrak{S}_\lambda} \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(1)}\right) < \cdots < f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(\mu)}\right)\right\}$$
$$\times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(\mu)}\right) < f_*\left(z+\sqrt{\mathbf{\Sigma}}U^{\sigma(k)}\right)\right\} \times \left(U^{\sigma(1)}, \ldots, U^{\sigma(\mu)}\right).$$

Let $\eta \colon \mathbb{R}^{d\mu} \to \mathbb{R}_+$ be a smooth map with compact support. We have

$$\mathbb{E}\left[\eta(V)\right] = \frac{1}{(\lambda-\mu)!} \sum_{\sigma \in \mathfrak{S}_\lambda} \int \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(1)}\right) < \cdots < f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(\mu)}\right)\right\}$$
$$\times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(\mu)}\right) < f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(k)}\right)\right\}$$
$$\times \eta\left(u_{\sigma(1)}, \ldots, u_{\sigma(\mu)}\right) p_U^d(u_1)\ldots p_U^d(u_\lambda)\mathrm{d}u_1\ldots\mathrm{d}u_\lambda.$$

However, observe that, for each $k = \mu+1, \ldots, \lambda$, we have

$$\int \mathbb{1}\left\{f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(\mu)}\right) < f_*\left(z+\sqrt{\mathbf{\Sigma}}u_{\sigma(k)}\right)\right\} p_U^d(u_{\sigma(k)})\mathrm{d}u_{\sigma(k)} = 1 - Q_{z,\mathbf{\Sigma}}^{f_*}\left(u_{\sigma(\mu)}\right).$$

We deduce the desired result. Since the composition of lower semicontinuous functions is lower semicontinuous and since f is continuous, when R is continuous, the function $((z, p, q, \hat{\mathbf{\Sigma}}, r), v) \mapsto p_{z,\hat{\mathbf{\Sigma}}/R(\hat{\mathbf{\Sigma}})}(v)$ is lower semicontinuous. □

Furthermore, under assumptions detailed in Section 2.2, we verify that **H**1 and **H**2 hold.

Proposition 5.4. *Suppose that the objective function f satisfies* **F**1-**F**2*, that the normalization function R satisfies* **R**1-**R**2*, and that the stepsize change Γ is such that* **Γ**1 *hold. Suppose moreover that ρ satisfies* **ρ**1-**ρ**2*. Consider the Markov chain $\Theta = \{(z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)\}_{t\in\mathbb{N}}$ defined by (5.3). Define the functions F_Θ and α_Θ via (5.10) and (5.4) respectively. Then, Θ follows (5.16), and* **H**1-**H**2 *hold.*

Proof. By Lemma 5.2, we find that **H**1 holds , with ζ_{V} being the Lebesgue measure on $\mathsf{V} = \mathbb{R}^{d\mu}$. Furthermore, using **R**2, **ρ**2 and **Γ**1, we deduce, by composition, that **H**2 is satisfied. □

5.2 Finding steadily attracting states

In this section and in Section 5.3, we prove that the control model (5.16) satisfies condition **H**3.[9] This is required to apply Theorem 4.1[10] and find that the Markov chain Θ obeying to (5.3) is an irreducible aperiodic T-chain. In this section, we focus on the existence of steadily attracting states. This is formalized in the next proposition.

Proposition 5.5. *Suppose that the objective function f satisfies* **F**1-**F**2*, that the stepsize change satisfies* **Γ**1-**Γ**2*, that the normalization functions R and ρ satisfy* **R**1-**R**2 *and* **ρ**1-**ρ**2 *respectively, and that the sampling distribution ν_U^d is such that* **N**1 *holds.*

Then $(0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ is a steadily attracting state for the control model (5.16) with the functions F_Θ and α_Θ given by (5.10) and (5.4), respectively.

Proof. Let $\theta_0 \in \mathsf{X}$. By Lemma 5.3, we find v_1 such that $S^1_{\theta_0}(v_1) = (0, p_1, q_1, \hat{\mathbf{\Sigma}}_1, r_1)$. If $q_1 \neq 0$, by Lemma 5.5, we set v_2, v_3 such that $S^3_{\theta_0}(v_{1:3}) = (0, p_3, 0, \hat{\mathbf{\Sigma}}_3, r_3)$. Using Lemma 5.6 we reach via a $4(d-1)$ steps a state $\theta = (0, \cdot, 0, \mathbf{I}_d, \cdot)$. Using Lemma 5.7, we complete the path v_1 in case $q_1 = 0$ or v_1, v_2, v_3 otherwise into $v_{1:\infty} = (v_1, v_2, \dots) \in \overline{\mathcal{O}^\infty_{\theta_0}}$ such that $\lim_{k\to\infty} S^k_{\theta_0}(v_{1:k}) = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$. This implies [19, Corollary 4.5] that $(0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ is a steadily attractive state. □

The proof of Proposition 5.5 relies on Lemmas 5.3 and 5.5 to 5.7 below. First, we state the next proposition, which is useful to provide candidates for the paths between an initial state and the steadily attracting state given by Proposition 5.5.

Proposition 5.6. *In the context of Proposition 5.5, let $\theta_0 = (z_0, p_0, q_0, \hat{\mathbf{\Sigma}}_0, r_0) \in \mathsf{X}$, let $k \geqslant 1$ and $v_{1:k} = (v_1, \dots, v_k) \in \mathsf{V}^k$ be such that for $i = 1, \dots, k$, we have $v_i = [v_i^1, \dots, v_i^\mu] \in \mathbb{R}^{d\mu}$ with $v_i^1 = \cdots = v_i^\mu \in \mathbb{R}^d$. Then, $v_{1:k} \in \overline{\mathcal{O}^k_{\theta_0}}$.*

Proof. We prove here that $v_1 = [\bar{v}_1, \dots, \bar{v}_1] \in \overline{\mathcal{O}^1_{\theta_0}}$. By Lemma 5.2, it is sufficient to prove that there exists a sequence $\{w_n = [w_n^1, \dots, w_n^\mu] \in \mathbb{R}^{d\mu}\}_{n\in\mathbb{N}}$ which converges to v_1 such that $p_{z_0,\mathbf{\Sigma}_0}(\mathbf{\Sigma}_0^{-1/2} w_n) > 0$ for all $n \in \mathbb{N}$, where $\mathbf{\Sigma}_0 = R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0$ and $p_{z_0,\mathbf{\Sigma}_0}$ is the density defined via (5.17). Moreover, by **N**1 and by definition of $p_{z_0,\mathbf{\Sigma}_0}$, it is sufficient to prove that for every $n \in \mathbb{N}$, $f(z_0 + w_n^1) < \cdots < f(z_0 + w_n^\mu)$. Furthermore, by **F**1, for every $n \in \mathbb{N}$, there exists $z_n^1, \dots, z_n^\mu \in \mathsf{B}(z_0 + \bar{v}_1, 1/n)$ such that $f(z_n^1) < \cdots < f(z_n^\mu)$. We take $w_n^i = z_n^i - z_0$ for $i = 1, \dots, \mu$ and $n \in \mathbb{N}$. Then w_n converges to v_1 and belongs to $\mathcal{O}^1_{\theta_0}$, so that $v_1 \in \overline{\mathcal{O}^1_{\theta_0}}$. Similarly, $v_2 \in \overline{\mathcal{O}^1_{S^1_{\theta_0}(x_1)}}$ for all x_1 and using the continuity of $v \mapsto S^1_{\theta_0}(v)$ in v_1, we deduce that $v_{1:2} \in \overline{\mathcal{O}^2_{\theta_0}} = \overline{\{(x_1, x_2) | p^1_{\theta_0}(x_1) \times p_{S^1_{\theta_0}(x_1)}(x_2) > 0\}}$. Similarly we obtain that $v_{1:k} \in \overline{\mathcal{O}^k_{\theta_0}}$. □

The following lemma is the first step to build a path between an arbitrary initial state $\theta_0 \in \mathsf{X}$ to the steadily attracting state $\theta^* = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ given in Proposition 5.5. More precisely, it shows that from $\theta_0 \in \mathsf{X}$, we can reach via a one-step path a state θ_1 such that $z_1 = 0$.

Lemma 5.3. *In the context of Proposition 5.5, let $\theta_0 = (z_0, p_0, q_0, \hat{\mathbf{\Sigma}}_0, r_0) \in \mathsf{X}$. Then there exists $\theta_1 = (0, p_1, q_1, \hat{\mathbf{\Sigma}}_1, r_1) \in \mathsf{X}$ and $v_1 \in \overline{\mathcal{O}^1_{\theta_0}}$ such that $S^1_{\theta_0}(v_1) = \theta_1$. Moreover, we can choose v_1 as a function of z_0 such that v_1 goes to 0 when z_0 tends to 0.*

Proof. Let $v_1 = -c_m^{-1} \times [z_0, \dots, z_0] \in (\mathbb{R}^d)^\mu$. It belongs to $\overline{\mathcal{O}^1_{\theta_0}}$ by Proposition 5.6. Set $\theta_1 = (z_1, p_1, q_1, \hat{\mathbf{\Sigma}}_1, r_1) = S^1_{\theta_0}(v_1)$. Then, $z_1 = F_z(z_0, p_0, q_0, \hat{\mathbf{\Sigma}}_0 / R(\hat{\mathbf{\Sigma}}_0), r_0; v_1) = r_1^{-1/2}\Gamma(p_1)^{-1} \times (z_0 - c_m \times c_m^{-1}\mathbf{w}_m^\top [z_0, \dots, z_0]) = 0$, see (5.5). We have used in particular that $\sum w_i^m = 1$. □

[9] Or condition **H**4 if we assume no cumulation on the stepsize or the covariance matrix.
[10] Or Theorem 4.2.

We make the following observation when the mean z_0 is in 0: by performing one step via $v_1 = (u_1, \dots, u_1)$ for any u_1 in $\mathbb{R}^d$, we can find a zero mean again in two steps by choosing a path $v_{1:2}$ appropriately.

Lemma 5.4. *In the context of Proposition 5.5, let $\theta_0 = (0, p_0, q_0, \hat{\mathbf{\Sigma}}_0, r_0) \in \mathsf{X}$. Then, given $v_1 = (u_1, \dots, u_1) \in \overline{\mathcal{O}^1_{\theta_0}}$ for some $u_1 \in \mathbb{R}^d$, and by defining $\theta_1 = (z_1, p_1, q_1, \hat{\mathbf{\Sigma}}_1, r_1) = S^1_{\theta_0}(v_1)$ and $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1$, we have that $v_{1:2} = [v_1, v_2] \in \overline{\mathcal{O}^2_{\theta_0}}$ and $\theta_2 = (z_2, p_2, q_2, \hat{\mathbf{\Sigma}}_2, r_2) = S^2_{\theta_0}(v_{1:2})$ satisfies $z_2 = 0$.*

Proof. By Proposition 5.6, we have $v_1 \in \overline{\mathcal{O}^1_{\theta_0}}$ and $v_{1:2} = [v_1, v_2] \in \overline{\mathcal{O}^2_{\theta_0}}$. Moreover, we have

$$z_2 = r_2^{-1/2}\Gamma(p_2)^{-1} \times \left(r_1^{-1/2}\Gamma(p_1)^{-1} \times (0 + c_m u_1) - c_m r_1^{-1/2}\Gamma(p_1)^{-1}u_1 \right) = 0$$

ending the proof. □

Next, from any initial state $\theta_0 \in \mathsf{X}$ with $z_0 = 0$, we reach via a two-steps path a state $\theta_2 \in \mathsf{X}$ with $z_2 = 0$ and $q_2 = 0$.

Lemma 5.5. *In the context of Proposition 5.5, let $\theta_0 = (0, p_0, q_0, \hat{\mathbf{\Sigma}}_0, r_0) \in \mathsf{X}$ such that $q_0 \neq 0$. Then, there exist $\theta_2 = (0, p_2, 0, \hat{\mathbf{\Sigma}}_2, r_2) \in \mathsf{X}$ and $v_{1:2} \in \overline{\mathcal{O}^2_{\theta_0}}$ such that $S^2_{\theta_0}(v_{1:2}) = \theta_2$. Moreover, we can choose $v_{1:2}$ such that $v_{1:2} \to 0$ when q_0 tends to 0.*

Proof. Let $u_1 \in \mathbb{R}^d$ and set $v_1 = (u_1, \dots, u_1)$. It belongs to $\overline{\mathcal{O}^1_{\theta_0}}$ by Proposition 5.6. Then, define

$$\theta_1 = (z_1, p_1, q_1, \hat{\mathbf{\Sigma}}_1, r_1) = F_\Theta\left(\theta_0, \alpha_\Theta\left(\theta_0, v_1\right)\right) = S^1_{\theta_0}(v_1) \ .$$

Then, define $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1 \in \overline{\mathcal{O}^1_{\theta_1}}$, and

$$\theta_2 = (z_2, p_2, q_2, \hat{\mathbf{\Sigma}}_2, r_2) = F_\Theta\left(\theta_1, \alpha_\Theta\left(\theta_1, v_2\right)\right) = S^2_{\theta_0}(v_{1:2}) \ .$$

Then, by Lemma 5.4, $v_{1:2} = (v_1, v_2) \in \overline{\mathcal{O}^2_{\theta_0}}$ and $z_2 = 0$. Moreover,

$$\begin{aligned} q_2 &= (1-c_c)^2 r_0^{-1/2} r_1^{-1/2} q_0 + (1-c_c) r_1^{-1/2}\sqrt{c_c(2-c_c)\mu_{\text{eff}}}u_1 \\ &\quad - r_1^{-1/2}\Gamma(p_1)^{-1}\sqrt{c_c(2-c_c)\mu_{\text{eff}}}u_1 \\ &= r_1^{-1/2} \times \left[(1-c_c)^2(r_0^{-1/2}q_0 + \left(1 - c_c - \Gamma(p_1)^{-1}\right) \times \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \times u_1\right] \ . \end{aligned}$$

Let $\kappa \in \mathbb{R}$, and choose $u_1 = \kappa q_0$. Since $v \mapsto S^2_{\theta_0}(v)$ is continuous, then both r_1 and q_2 depend continuously on κ. Moreover, we have

$$q_2 = r_1^{-1/2} \times \left[(1-c_c)^2 r_0^{-1/2} + \left(1 - c_c - \Gamma(p_1)^{-1}\right) \times \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\kappa\right] \times q_0 \ .$$

But, as $r_1 > 0$, and as $\Gamma(p_1)^{-1} = \Gamma\left((1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}R(\hat{\mathbf{\Sigma}}_0)}\kappa\hat{\mathbf{\Sigma}}_0^{-1/2}q_0\right)^{-1}$ is less that $1 - c_c$ when $\kappa \to \pm\infty$ by **Γ**2, then by the intermediate value theorem (since Γ is continuous by **Γ**1), there exists $\kappa \in \mathbb{R}$ such that $q_2 = 0$. With the above choice of $u_1 = \kappa q_0$, $v_1 = (u_1, \dots, u_1)$ and $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1 \in \overline{\mathcal{O}^1_{\theta_1}}$, we see that $v_{1:2} \to 0$ when q_0 tends to 0. □

From an initial state θ_0 with $z_0 = q_0 = 0$, we reach via a $4(d-1)$-steps path a state $\theta_{4(d-1)}$ with $z_{4(d-1)} = q_{4(d-1)} = 0$, and $\hat{\mathbf{\Sigma}}_{4(d-1)} = \mathbf{I}_d$. This is achieved by applying $(d-1)$ times the following lemma successively to the k-th largest (counted with multiplicity) eigenvalue of $\hat{\mathbf{\Sigma}}_0$, for $k = 2, \dots, d$. For the sake of conciseness, the proof of Lemma 5.6 is delayed to Appendix A.

Lemma 5.6. *In the context of Proposition 5.5, let $\theta_0 = (0, p_0, 0, \hat{\mathbf{\Sigma}}_0, r_0)$. Consider an orthonormal basis $\mathcal{B}$ of $\mathbb{R}^d$ composed of eigenvectors of $\hat{\mathbf{\Sigma}}_0$ such that the matrix $\hat{\mathbf{\Sigma}}_0$ writes in the basis $\mathcal{B}$ as*

$$[\hat{\mathbf{\Sigma}}_0]_{\mathcal{B}} = \operatorname{diag}\left(\lambda_1, \dots, \lambda_d\right),$$

with $\lambda_1 = \lambda_2 = \cdots = \lambda_{k-1} \geqslant \lambda_k \geqslant \cdots \geqslant \lambda_d$ for some $2 \leqslant k \leqslant d$. Then, there exists $\gamma > 0$, such that the matrix $\hat{\mathbf{\Sigma}}_4$ defined by

$$[\hat{\mathbf{\Sigma}}_4]_{\mathcal{B}} = \gamma \times \operatorname{diag}\left(\lambda_1, \dots, \lambda_{k-1}, \lambda_{k-1}, \lambda_{k+1}, \dots, \lambda_d\right), \tag{5.19}$$

is such that for some $p_4 \in \mathbb{R}^d$ and $r_4 > 0$, and $v_{1:4} \in \overline{\mathcal{O}^4_{\theta_0}}$, we have $S^4_{\theta_0}(v_{1:4}) = \theta_4 = (0, p_4, 0, \hat{\mathbf{\Sigma}}_4, r_4)$.

Finally, as stated in the next lemma, from an initial state $\theta_0 \in \mathsf{X}$ such that $z_0 = q_0 = 0$ and $\hat{\mathbf{\Sigma}}_0 = \mathbf{I}_d$, we can reach any neighborhood of the state $\theta^* = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$.

Lemma 5.7. *In the context of Proposition 5.5, let $\theta_0 = (0, p_0, 0, \mathbf{I}_d, r_0) \in \mathsf{X}$. Then there exists $v_{1:\infty} \in \overline{\mathcal{O}^\infty_{\theta_0}}$ such that $\lim S^t_{\theta_0}(v_{1:t}) = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ when $t \to \infty$.*

Proof. Define $v_{1:\infty}$, by $v_t = 0 \in \mathbb{R}^{d\times\mu}$ for all $t \geqslant 1$. By Proposition 5.6, we have $v_{1:\infty} \in \overline{\mathcal{O}^\infty_{\theta_0}}$. Denote

$$\theta_{t+1} = (z_{t+1}, p_{t+1}, q_{t+1}, \hat{\mathbf{\Sigma}}_{t+1}, r_{t+1}) = F_\Theta(\theta_t, \alpha_\Theta(\theta_t, v_{t+1})).$$

Since, $\theta_0 = (0, p_0, 0, \mathbf{I}_d, r_0)$ and $v_t = 0$, by induction, we have $\hat{\mathbf{\Sigma}}_{t+1} = \mathbf{I}_d$, $z_{t+1} = 0$, $q_{t+1} = 0$ $r_{t+1} = R((1 - c_1 - c_\mu)\mathbf{I}_d) = 1 - c_1 - c_\mu$ and $p_{t+1} = (1 - c_\sigma)p_t$. Since $0 \leqslant 1 - c_\sigma < 1$, then $(z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)$ tends to $(0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ when $t \to \infty$, ending the proof. □

Lastly, as a consequence of Proposition 5.5, we prove that given any normalized covariance matrix $\hat{\mathbf{\Sigma}}^* \in \mathcal{S}^d_{++}$ such that $\rho(\hat{\mathbf{\Sigma}}^*) = 1$, we can find a value for the path $p^* \in \mathbb{R}^d$, for the variable $r^* > 0$, such that the state $\theta^* = (0, p^*, 0, \hat{\mathbf{\Sigma}}^*, r^*) \in \mathsf{X}$ with normalized mean and normalized path for the rank-one update equal to zero is steadily attracting. In Section 5.3, we use these steadily attracting states to prove the controllability condition stated in Proposition 5.7.

Corollary 5.1. *Consider the context of Proposition 5.5. Let $\hat{\mathbf{\Sigma}}^* \in \mathcal{S}^d_{++}$ be such that $\rho(\hat{\mathbf{\Sigma}}^*) = 1$. Then, there exist $p^* \in \mathbb{R}^d$ and $r^* > 0$ such that $\theta^* = (0, p^*, 0, \hat{\mathbf{\Sigma}}^*, r^*) \in \mathsf{X}$ is a steadily attracting state.*

Proof. By Proposition 5.5, we know that $\theta_0 = (0, 0, 0, \mathbf{I}_d, 1 - c_1 - c_\mu)$ is a steadily attracting state. Hence, in order to prove that a state $\theta^* \in \mathsf{X}$ is steadily attracting, it is sufficient, as explained below, to prove

(i) that there exist $k \in \mathbb{N}$ and $v_{1:k} \in \overline{\mathcal{O}^k_{\theta_0}}$ such that $S^k_{\theta_0}(v_{1:k}) = \theta^*$.

Indeed, assume we have proven (i) and let V be a neighborhood of θ^* and let $\theta \in \mathsf{X}$. Then, by continuity of $w_{1:k} \mapsto S^k_{\theta_0}(w_{1:k})$ around $v_{1:k}$, there exists $v^*_{1:k} \in \mathcal{O}^k_{\theta_0}$ such that $S^k_{\theta_0}(v^*_{1:k}) \in V$. Since $x \mapsto p^k_x(v^*_{1:k})$ is lower semicontinuousand $x \mapsto S^k_x(v^*_{1:k})$ is continuous, then there exists a neighborhood U of θ_0 such that for every $x \in U$, $p^k_x(v^*_{1:k}) > 0$, i.e., $v^*_{1:k} \in \mathcal{O}^k_x$, and $S^k_x(v^*_{1:k}) \in V$. Moreover, since θ_0 is steadily attracting, then there exists $T > 0$ such that for every $t \geqslant T$, there exists $w_{1:t} \in \mathcal{O}^k_\theta$ such that $S^t_\theta(w_{1:t}) \in U$, hence $[w_{1:t}, v^*_{1:k}] \in \mathcal{O}^{t+k}_\theta$ and $S^{t+k}_\theta([w_{1:t}, v^*_{1:k}]) \in V$ and hence θ^* is a steadily attracting state.

Let $\hat{\mathbf{\Sigma}}^* \in \mathcal{S}^d_{++}$ be such that $\rho(\hat{\mathbf{\Sigma}}^*) = 1$. We proceed now as in Lemma 5.6 to prove (i) for a state θ^* that is equal to $(0, p^*, 0, \hat{\mathbf{\Sigma}}^*, r^*)$ for p^* and r^* constructed below. For $i = 1, \dots, d$, let λ_i be the i-th largest eigenvalue of $\hat{\mathbf{\Sigma}}^*$ (counted with multiplicity), and $(e_1, \dots, e_d)$ an orthonormal basis of eigenvectors of $\hat{\mathbf{\Sigma}}^*$ such that $\hat{\mathbf{\Sigma}}^* e_i = \lambda_i e_i$. Then, let κ and κ' be real numbers, and by Proposition 5.6, define $v_{1:4} \in \overline{\mathcal{O}^4_{\theta_0}}$ by

$$v_1 = \kappa[e_1, \dots, e_1] \in \mathbb{R}^{d\mu},\ v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1,\ v_3 = \kappa'[e_1, \dots, e_1],\ v_4 = -r_3^{-1/2}\Gamma(p_3)^{-1}v_3,$$

where $\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S^t_{\theta_0}(v_{1:t})$ for $t = 1, 2, 3, 4$. Then, as in the proof of Lemma 5.6, there exist values of κ and κ' in $\mathbb{R}$ such that $z_4 = q_4 = 0$ and such that there exists $\rho_4 > 0$ with $\hat{\mathbf{\Sigma}}_4$ satisfying $\hat{\mathbf{\Sigma}}_4 e_1 = \rho_4(1 - c_1 - c_\mu)^{-4(d-1)}\lambda_1 e_1$ and $\hat{\mathbf{\Sigma}}_4 e_k = \rho_4(1 - c_1 - c_\mu)^4 e_k$ for $k = 2, \dots, d$.

Then, by repeating these steps with $e_2, \dots, e_d$ instead of e_1 and $\lambda_2, \dots, \lambda_d$ instead of λ_1, then there exist $v_{1:4d} \in \overline{\mathcal{O}^{4d}_{\theta_0}}$ and $\rho_{4d} > 0$ such that $\theta_{4d} = (z_{4d}, p_{4d}, q_{4d}, \hat{\mathbf{\Sigma}}_{4d}, r_{4d}) = S^{4d}_{\theta_0}(v_{1:4d})$ satisfies $z_{4d} = q_{4d} = 0$ and $\hat{\mathbf{\Sigma}}_{4d} e_k = \rho_{4d}\lambda_k e_k$ for $k = 1, \dots, d$. Hence $\hat{\mathbf{\Sigma}}_{4d} = \rho_{4d}\hat{\mathbf{\Sigma}}^*$. But since $\rho(\hat{\mathbf{\Sigma}}^*) = 1$ and $\rho(\hat{\mathbf{\Sigma}}_{4d}) = 1$, then $\rho_{4d} = 1$, i.e. $\hat{\mathbf{\Sigma}}_{4d} = \hat{\mathbf{\Sigma}}^*$ such that we have proven (i) for $\theta^* = (0, p_{4d}, 0, \hat{\mathbf{\Sigma}}^*, r_{4d})$ and in turn that $\theta^* = (0, p_{4d}, 0, \hat{\mathbf{\Sigma}}^*, r_{4d})$ is a steadily attracting state. □

5.3 Controllability condition

In the previous section, we prove that the control model (5.16) admits steadily attracting states. In the current section, we prove that a controllability condition, as required to satisfy the assumptions **H**3 or **H**4, is satisfied at a steadily attracting state. By combining Corollary 5.1 and the following Proposition 5.7, we prove **H**3 or **H**4. For a finite-dimensional vectorial space E equipped with a norm $\|\cdot\|$ and an element $h \in \mathsf{E}$, we use the notation $o(h)$, respectively $O(h)$, to be understood as $o(\|h\|)$, respectively $O(\|h\|)$. Besides, it does not depend on the chosen norm, since all norms on a finite-dimensional space induce the same topology.

Proposition 5.7. *Suppose that the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν_U^d satisfy* **F**1-**F**2, **R**1-**R**3, **ρ**1-**ρ**2, **Γ**1-**Γ**3 *and* **N**1, *respectively.*

Consider the control model (5.16) with the functions F_Θ and α_Θ defined by (5.10) and (5.4) respectively.

Then, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T > 0$ and $v_{1:T} \in \overline{\mathcal{O}_{\theta_0}^T}$ such that $S_{\theta_0}^T$ is differentiable at $v_{1:T}$, and, by denoting $(z_T, p_T, q_T, \hat{\mathbf{\Sigma}}_T, r_T) = S_{\theta_0}^T(v_{1:T})$, we have

(a) if $c_c \neq 1$, $c_\sigma \neq 1$, $1 - c_c \neq (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$ and $c_\mu > 0$, then $\mathcal{D}S_{\theta_0}^T(v_{1:T})$ is of maximal rank;

(b) if $c_c = 1$ and $c_\sigma \neq 1$, then, for every $(z, p, \mathbf{\Sigma}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, there exist $q \in \mathbb{R}^d$ and $r \in \mathbb{R}$ such that $(z, p, q, \mathbf{\Sigma}, r) \in \mathrm{rge}\ \mathcal{D}S_{\theta_0}^T(v_{1:T})$;

(c) if $c_c \neq 1$, $c_\sigma = 1$ and $c_\mu > 0$, then, for every $(z, q, \mathbf{\Sigma}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\}) \times \mathbb{R}$, there exists $p \in \mathbb{R}^d$ such that $(z, p, q, \mathbf{\Sigma}, r) \in \mathrm{rge}\ \mathcal{D}S_{\theta_0}^T(v_{1:T})$;

(d) if $c_c = c_\sigma = 1$, then, for every $(z, \mathbf{\Sigma}) \in \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, there exist $(p, q) \in \mathbb{R}^d \times \mathbb{R}^d$ and $r \in \mathbb{R}$, such that $(z, p, q, \mathbf{\Sigma}, r) \in \mathrm{rge}\ \mathcal{D}S_{\theta_0}^T(v_{1:T})$.

Before proving Proposition 5.7, we state the two following lemmas, which characterize the derivatives of the normalization function ρ and of the transition map $S_{\theta_0}^1$, respectively.

Lemma 5.8. *Consider a positively homogeneous function $R\colon \mathcal{S}_{++}^d \to \mathbb{R}_{++}$. Let $\mathbf{A} \in \mathcal{S}_{++}^d$ and $\gamma > 0$ and suppose that R is differentiable at $\mathbf{A}$. Then, R is differentiable at $\gamma\mathbf{A}$ and $\mathcal{D}R(\gamma\mathbf{A}) = \mathcal{D}R(\mathbf{A})$.*

Proof. By Taylor expansion, we have, when $\mathbf{H} \in \mathcal{S}^d$ tends to 0, that

$$\begin{aligned} R(\gamma\mathbf{A} + \mathbf{H}) &= \gamma \times R(\mathbf{A} + \gamma^{-1}\mathbf{H}) = \gamma R(\mathbf{A}) + \gamma\mathcal{D}R(\mathbf{A})\gamma^{-1}\mathbf{H} + o(\mathbf{H}) \\ &= R(\gamma\mathbf{A}) + \mathcal{D}R(\mathbf{A})\mathbf{H} + o(\mathbf{H}) \end{aligned}$$

and thus, by Taylor expansion, R is differentiable at $\gamma\mathbf{A}$ and $\mathcal{D}R(\gamma\mathbf{A}) = \mathcal{D}R(\mathbf{A})$. □

Lemma 5.9. *Suppose* **Γ**1, **R**1, **R**2 *and* **ρ**2. *Let $\theta_0 = (z_0, p_0, q_0, \hat{\mathbf{\Sigma}}_0, r_0) \in \mathsf{X}$ and $v_1 \in \overline{\mathcal{O}_{\theta_0}^1}$. Then, if (a) $z_0 = q_0 = 0$ and $v_1 = 0$, or if (b) $p_1 = F_p(p_0, R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0; v_1) \neq 0$, and if moreover R is differentiable in $\hat{\mathbf{\Sigma}}_1 = F_{\mathbf{\Sigma}}(q_0, R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0, r_0; v_1)$ (see (5.14) for the definition of $F_{\mathbf{\Sigma}}$), then $v \in \mathsf{V} \mapsto S_{\theta_0}^1(v)$ is differentiable at v_1.*

Proof. Suppose (a). Then, for $h_1 = (h_1^1, \dots, h_1^\mu) \in \mathsf{V}$, using the update equations (5.5), (5.6), (5.7), (5.8), (5.9) we have, when $h_1 \to 0$,

$$S_{\theta_0}^1(v_1 + h_1) = \begin{pmatrix} \frac{R((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + o(\|h_1\|))^{-1/2}(0 + c_m\mathbf{w}_m^\top h_1)}{\Gamma((1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\mathrm{eff}}}R(\hat{\mathbf{\Sigma}}_0)^{1/2}\hat{\mathbf{\Sigma}}_0^{-1/2}\mathbf{w}_m^\top h_1)} \\ (1 - c_\sigma)p_0 + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\mathrm{eff}}}R(\hat{\mathbf{\Sigma}}_0)^{1/2}\hat{\mathbf{\Sigma}}_0^{-1/2}\mathbf{w}_m^\top h_1 \\ 0 + \sqrt{c_c(2 - c_v)\mu_{\mathrm{eff}}}\mathbf{w}_m^\top h_1 \\ \frac{(1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + o(\|h_1\|)}{\rho((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + o(\|h_1\|))} \\ R((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + o(\|h_1\|)) \end{pmatrix} .$$

However, by **Γ**1, the stepsize change Γ is locally Lispchitz, hence

$$\Gamma((1 - c_\sigma)p_0 + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\mathrm{eff}}}R(\hat{\mathbf{\Sigma}}_0)^{1/2}\hat{\mathbf{\Sigma}}_0^{-1/2}\mathbf{w}_m^\top h_1) = \Gamma((1 - c_\sigma)p_0) + O(\|h_1\|) = \Gamma((1 - c_\sigma)p_0) + o(1) \ .$$

Moreover, by **ρ**2, ρ is differentiable at $(1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0$. Hence by Taylor expansion

$$\rho((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + o(\|h_1\|)) = \rho((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0) + o(\|h_1\|) \ .$$

Likewise, by assumption, R is differentiable at $(1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0$. Thus,

$$R((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + o(\|h_1\|)) = R((1 - c_1 - c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0) + o(\|h_1\|) \ .$$

Therefore,

$$S^1_{\theta_0}(v_1+h_1) = S^1_{\theta_0}(v_1) + \begin{pmatrix} R((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0)^{-1/2}\Gamma((1-c_\sigma)p_0)^{-1}c_m\mathbf{w}_m^\top h_1 \\ \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\mathbf{\Sigma}}_0)^{1/2}\hat{\mathbf{\Sigma}}_0^{-1/2}\mathbf{w}_m^\top h_1 \\ \sqrt{c_c(2-c_v)\mu_{\text{eff}}}\mathbf{w}_m^\top h_1 \\ 0 \\ 0 \end{pmatrix} + o(\|h_1\|) \ ,$$

which proves by Taylor expansion that $S^1_{\theta_0}$ is differentiable at $v_1 = 0$.

Now, suppose (b). By $\mathbf{\Gamma}1$, Γ is differentiable at p_1, by $\boldsymbol{\rho}2$, ρ is differentiable on $\mathcal{S}^d_{++}$, and by assumption, R is differentiable at $\hat{\mathbf{\Sigma}}_1 = \mathbf{A}_1/\rho(\mathbf{A}_1)$ where $\mathbf{A}_1 = (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0 + c_1 q_1(q_1)^\top + c_\mu \sum_{i=1}^\mu w_i^c v_1^i (v_1^i)^\top$. Since R is positively homogeneous, it is also differentiable in any multiple by a scalar of $\hat{\mathbf{\Sigma}}_1$, so in $\mathbf{A}_1$. Thus, by composition $S^1_{\theta_0}$ is differentiable at v_1. □

We prove now Proposition 5.7. The first step of the proof consists in the following lemma which applies to all cases (a)-(d) in Proposition 5.7. It provides a path $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$ (where θ_0 is the steadily attracting state found in Section 5.2) such that the range of $\mathcal{D}S^T_{\theta_0}$ covers all elements in the tangent space relative to the covariance matrix variable. The proof of Lemma 5.10 is delayed to Appendix B.

Lemma 5.10. *Suppose that the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν_U^d satisfy* **F**1-**F**2, **R**1-**R**3, $\boldsymbol{\rho}$1-$\boldsymbol{\rho}$2, $\mathbf{\Gamma}$1-$\mathbf{\Gamma}$3 *and* **N**1*, respectively. Consider the control model (5.16) with the functions F_Θ and α_Θ defined by (5.10) and (5.4) respectively.*

Then, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T \in \mathbb{N}$, $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$, and W a subspace of V^T, such that:

(i) $S^T_{\theta_0}$ is differentiable at $v_{1:T}$;

(ii) for every $h_{\mathbf{\Sigma}} \in \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, there exists $h_z, h_p \in \mathbb{R}^d$, $h_r \in \mathbb{R}$, and $h_{1:T} \in \mathsf{W}$ such that $\mathcal{D}S^T_{\theta_0}(v_{1:T})h_{1:T} = [h_z, h_p, 0, h_{\mathbf{\Sigma}}, h_r]$;

(iii) $z_T = q_T = 0$ and $p_T \neq 0$;

where $\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S^t_{\theta_0}(v_{1:t})$ for $t = 1, \dots, T$.

The next lemma is the second step of the proof of Proposition 5.7. It deduces from Lemma 5.10 a path in which the transition map is differentiable and is of interest to apply Theorem 4.1 or Theorem 4.2. It applies to all cases (a)-(d). We delay once more the proof of Lemma 5.11 to Appendix B.

Lemma 5.11. *Suppose that the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν_U^d satisfy* **F**1-**F**2, **R**1-**R**3, $\boldsymbol{\rho}$1-$\boldsymbol{\rho}$2, $\mathbf{\Gamma}$1-$\mathbf{\Gamma}$3 *and* **N**1*, respectively. Consider the control model (5.16) with the functions F_Θ and α_Θ defined by (5.10) and (5.4) respectively.*

Then, there exist $\theta_0 \in \mathsf{X}$ a steadily attracting state, and $p \in \mathbb{R}^d_{\neq 0}$, such that, for every $j \in \mathbb{N}$, there exist $T \in \mathbb{N}$ and $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$, with $S^T_{\theta_0}$ being differentiable at $v_{1:T}$, and

$$S^T_{\theta_0}(v_{1:T}+h_{1:T}) = S^T_{\theta_0}(v_{1:T}) + \mathbf{C}_j \times L(h_{1:T}) + o(h_{1:T}) \tag{5.20}$$

for every $h_{1:T} \in \mathsf{W}_L$, where W_L is a well-chosen subspace of V^T, $L\colon \mathsf{W}_L \to \mathbb{R}^{s-1} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ is a surjective linear map, $s = d(d+1)/2$, and $\mathbf{C}_j$ is a matrix of the form:

$$\mathbf{C}_j = \begin{bmatrix} * & \cdots & * & 0 & 0 & L^z \\ * & \cdots & * & \multicolumn{2}{c}{\multirow{3}{*}{$\mathbf{L}_j^{p,q}$}} & * \\ * & \cdots & * & & & * \\ \mathbf{L}_1^{\mathbf{\Sigma}} & \dots & \mathbf{L}_{s-1}^{\mathbf{\Sigma}} & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.21}$$

with $(\mathbf{L}_1^{\mathbf{\Sigma}}, \dots, \mathbf{L}_{s-1}^{\mathbf{\Sigma}})$ being a basis of $\ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_T)$ (with $\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S^t_{\theta_0}(v_{1:t})$ for $t = 0, \dots, T$), $L_z \in \mathbb{R}_{\neq 0}$, and

$$\mathbf{L}_j^{p,q} = \begin{bmatrix} (1-c_\sigma)^3 c_{j+1}^p R(\hat{\mathbf{\Sigma}}_T)^{1/2}\hat{\mathbf{\Sigma}}_T^{-1/2} & (1-c_\sigma)^1 c_{j+3}^p R(\hat{\mathbf{\Sigma}}_T)^{1/2}\hat{\mathbf{\Sigma}}_T^{-1/2} \\ (1-c_c)^3(1-c_1-c_\mu)^{-3/2} d_{j+1}^p \mathbf{I}_d & (1-c_c)^1(1-c_1-c_\mu)^{-1/2} d_{j+3}^p \mathbf{I}_d \end{bmatrix} , \tag{5.22}$$

where $c_k^p := (1-c_\sigma-(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^k p))^{-1}\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}$ and $d_k^p := (1-c_1-c_\mu)^{-1/2}\left[1-c_c-\Gamma((1-c_\sigma)^k p)^{-1}\right]\sqrt{c_c(2-c_c)\mu_{\text{eff}}}$. The symbol $*$ in (5.21) represents the elements of the matrix $\mathbf{C}_j$ that we do not give explicitly (their values do not change the rank of $\mathbf{C}_j$).

Next, in order to deduce the case (a) in Proposition 5.7 from Lemma 5.11, we first show in the next lemma that the matrix $\mathbf{L}_j^{p,q}$ defined via (5.22) is invertible when the integer j is sufficiently large.

Lemma 5.12. *In the context of Lemma 5.11, there exists $j \in \mathbb{N}$ such that, if $c_c \neq 1$, $c_\sigma \neq 1$, $1 - c_c \neq (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$, then the matrix $\mathbf{L}_j^{p,q}$ defined via (5.22) is invertible.*

Proof. We have, since $c_\sigma \neq 1$, $c_c \neq 1$:

$$\begin{bmatrix} (1-c_\sigma)^{-1} R(\hat{\mathbf{\Sigma}}_T)^{-1/2} \hat{\mathbf{\Sigma}}_T^{1/2} & 0 \\ 0 & (1-c_c)^{-1}(1-c_1-c_\mu)^{1/2}\mathbf{I}_d \end{bmatrix} \times \mathbf{L}_j^{p,q} \\ = \begin{bmatrix} (1-c_\sigma)^2 c_{j+1}^p \mathbf{I}_d & c_{j+3}^p \mathbf{I}_d \\ (1-c_c)^2(1-c_1-c_\mu)^{-1} d_{j+1}^p \mathbf{I}_d & d_{j+3}^p \mathbf{I}_d \end{bmatrix} .$$

Therefore, it is sufficient to find some $j \in \mathbb{N}$ such that the RHS in the above equation is invertible, i.e., such that the matrix

$$\mathbf{A}_j = \begin{bmatrix} (1-c_\sigma)^2 \left[1 - c_\sigma - (1-c_1-c_\mu)^{-1/2}\Gamma_{j+1}^{-1}\right] & 1 - c_\sigma - (1-c_1-c_\mu)^{-1/2}\Gamma_{j+3}^{-1} \\ (1-c_c)^2(1-c_1-c_\mu)^{-1}\left[1 - c_c - \Gamma_{j+1}^{-1}\right] & 1 - c_c - \Gamma_{j+3}^{-1} \end{bmatrix} ,$$

where $\Gamma_k = \Gamma((1-c_\sigma)^k p)$ for $k = j+2, j+4$, is full rank. Moreover, when $j \to \infty$, by continuity of Γ (by $\mathbf{\Gamma}1$), we have that Γ_{j+1} and Γ_{j+3} tend to $\Gamma(0)$. Hence,

$$\lim_{j\to\infty} \det \mathbf{A}_j = \begin{vmatrix} (1-c_\sigma)^2(1 - c_\sigma - (1-c_1-c_\mu)^{-1/2}\Gamma(0)^{-1}) & 1 - c_\sigma - (1-c_1-c_\mu)^{-1/2}\Gamma(0)^{-1} \\ (1-c_c)^2(1-c_1-c_\mu)^{-1}(1-c_c-\Gamma(0)^{-1}) & 1 - c_c - \Gamma(0)^{-1} \end{vmatrix} \\ = (1 - c_\sigma - (1-c_1-c_\mu)^{-1/2}\Gamma(0)^{-1}) \times (1 - c_c - \Gamma(0)^{-1}) \times \begin{vmatrix} (1-c_\sigma)^2 & 1 \\ (1-c_c)^2(1-c_1-c_\mu)^{-1} & 1 \end{vmatrix} ,$$

where $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ denotes the determinant of the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. However, $\Gamma(0)^{-1} > 1$ (by $\mathbf{\Gamma}3$) and $(1-c_1-c_\mu)^{-1} > 1$. Hence, there exists $j \in \mathbb{N}$, such that, if $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, then $\det \mathbf{L}_j^{p,q} \neq 0$. □

We can now end the proof of Proposition 5.7. Depending on the case (a)-(d), the end of the proof goes differently. We present here the proofs of cases (b) and (d), and we delay those of (a) and (d) to Appendix B.3.

Proof of Proposition 5.7(d). Suppose that $c_c = c_\sigma = 1$. Apply Lemma 5.11, we have then that the matrix $\mathbf{L}_j^{p,q}$ defined via (5.22) is the zero matrix. Then, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T > 0$ and $v_{1:T} \in \overline{\mathcal{O}_{\theta_0}^T}$ such that we have that $\operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T}) \supset \mathbb{R}^d \times \{0\} \times \{0\} \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\}) \times \{0\}$ and thus by taking $p = q = 0$ and $r = 0$, we have, for every $(z, \hat{\mathbf{\Sigma}}) \in \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, $(z, p, q, \hat{\mathbf{\Sigma}}, r) \in \operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T})$. □

Proof of Proposition 5.7(b). Suppose that $c_c = 1$ and $c_\sigma \neq 1$. By Lemma 5.11, there exist a steadily attracting state $\theta_0 \in \mathsf{X}$, $T > 0$ and $v_{1:T} \in \overline{\mathcal{O}_{\theta_0}^T}$ such that the matrix $\mathbf{L}_j^{p,q}$ defined via (5.22) satisfies:

$$\mathbf{L}_j^{p,q} = \begin{bmatrix} (1-c_\sigma)^3 c_j^p \mathbf{\Sigma}_T^{-1/2} & (1-c_\sigma) c_{j+2}^p \mathbf{\Sigma}_T^{-1/2} \\ 0 & 0 \end{bmatrix} , \tag{5.23}$$

with $\operatorname{rank} \mathbf{\Sigma}_T^{-1/2} = d$, and $c_j^p \neq 0$, $c_{j+2}^p \neq 0$. Thus, $\operatorname{rge} \mathbf{L}_j^{p,q} = \mathbb{R}^d \times \{0\}$ and $\operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T}) \supset \mathbb{R}^d \times \mathbb{R}^d \times \{0\} \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\}) \times \{0\}$, and thus by taking $q = 0$ and $r = 0$, we have, for every $(z, p, \hat{\mathbf{\Sigma}}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\mathbf{\Sigma}}_T}\rho^{-1}(\{1\})$, $(z, p, q, \hat{\mathbf{\Sigma}}, r) \in \operatorname{rge} \mathcal{D}S_{\theta_0}^T(v_{1:T})$. □

5.4 Proof of Theorem 3.1

In Sections 5.1 to 5.3, we have proven all required conditions to apply Theorem 4.1 or Theorem 4.2 to the Markov chain Θ defined in (5.3). The conclusion is summarized in the next theorem.

Theorem 5.1. *Suppose the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$, the normalization functions R and ρ, the stepsize change $\Gamma\colon \mathbb{R}^d \to \mathbb{R}_{++}$ and the sampling distribution ν_U^d satisfy* **F**1-**F**2, **R**1-**R**3, **ρ**1-**ρ**2, **Γ**1-**Γ**3 *and* **N**1, *respectively.*

Let $\Theta = \{(z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t)\}_{t\geqslant 1}$ be the normalized Markov chain associated to CMA-ES defined via (5.3) and P its transition kernel. Then,

(i) *if* $c_c, c_\sigma \in (0,1)$ *are such that* $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, *and if* $c_\mu > 0$, *then* P *is an irreducible aperiodic* T*-kernel, such that compact sets of* $\mathsf{X} = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$ *are small;*

(ii) *if* $c_c \in (0,1)$, $c_\sigma = 1$ *and* $c_\mu > 0$, *then the normalized chain* $\{(z_t, q_t, \hat{\boldsymbol{\Sigma}}_t, r_t)\}_{t\geqslant 1}$ *is a time-homogeneous Markov chain with an irreducible aperiodic* T*-kernel, such that compact sets of* $\mathsf{X}_2 = \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\}) \times \mathbb{R}_{++}$ *are small;*

(iii) *if* $c_\sigma \in (0,1)$ *and* $c_c = 1$, *then the normalized chain* $\{(z_t, p_t, \hat{\boldsymbol{\Sigma}}_t)\}_{t\geqslant 1}$ *is a time-homogeneous Markov chain with an irreducible aperiodic* T*-kernel, such that compact sets of* $\mathsf{X}_3 = \mathbb{R}^d \times \mathbb{R}^d \times \rho^{-1}(\{1\})$ *are small;*

(iv) *if* $c_c = c_\sigma = 1$, *then the normalized chain* $\{(z_t, \hat{\boldsymbol{\Sigma}}_t)\}_{t\geqslant 1}$ *is a time-homogeneous Markov chain with an irreducible aperiodic* T*-kernel, such that compact sets of* $\mathsf{X}_4 = \mathbb{R}^d \times \rho^{-1}(\{1\})$ *are small.*

Proof. By Proposition 5.3, the Markov chain Θ follows the control model (5.16), and by Proposition 5.4, **H**1 and **H**2 hold.

Suppose first that $c_c, c_\sigma \neq 1$ and $1 - c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$. By Proposition 5.7, there exist a steadily attracting state $\theta^* \in \mathsf{X}$, $T \geqslant 1$ $v_{1:T} \in \overline{\mathcal{O}^T_{\theta^*}}$ such that $\mathcal{D}S^T_{\theta^*}(v_{1:T})$ exists and is of maximal rank. Hence **H**3 holds, and we deduce then (i) by applying Theorem 4.1.

Now suppose that $c_c \neq 1$ and $c_\sigma = 1$, resp. $c_\sigma \neq 1$ and $c_c = 1$, and $c_c = c_\sigma = 1$. Then, by Corollary 4.1, $\Theta^q := \{(z_t, q_t, \hat{\boldsymbol{\Sigma}}_t, r_t)\}_{t\geqslant 1}$, resp. $\Theta^p := \{(z_t, p_t, \hat{\boldsymbol{\Sigma}}_t)\}_{t\geqslant 1}$, and $\Theta^r := \{(z_t, \hat{\boldsymbol{\Sigma}}_t)\}_{t\geqslant 1}$, defines a time-homogeneous Markov chain. Moreover, since θ^* is a steadily attracting state for Θ, then, by Proposition 5.7, resp. Θ^p, Θ^q and Θ^r, follows a control model which satisfies **H**4. Thus, by Theorem 4.2, we obtain (ii), (iii) and (iv). □

Our main result Theorem 3.1, stated in Section 2, is a consequence of Theorem 5.1 and of Theorem 4.3. Indeed, consider $\rho = \det(\cdot)^{1/d}$. It is a normalization function that satisfies $\boldsymbol{\rho}$1-$\boldsymbol{\rho}$2. By Proposition 5.2, the associated Markov chain Θ following (5.3) with this normalization function is a transformation of the chain Φ defined via (2.14) by the homeomorphism ξ defined in (5.1).

By Theorem 5.1, Θ is an irreducible aperiodic T-chain. By Theorem 4.3, so is Φ. Therefore [35, Theorem 6.2.5], compact sets are small sets.

6 Conclusion and perspectives

This paper expands a methodology to analyze irreducibility and other stability properties of complex Markov chains when they are expressed as nonsmooth state-space models. We apply the methodology in the context of optimization to the CMA-ES [22]. We prove irreducibility, aperiodicity and topological properties of a stochastic process obtained by normalizing the Markov chain that represents the state of CMA-ES when optimizing scaling-invariant functions. This is an important milestone to prove the linear convergence of CMA-ES.

Our stability analysis encompasses more general processes than the one underlying CMA-ES by considering an abstract stepsize change function. Compared to previous work [42], we relax the assumption on the stepsize change from $\mathcal{C}^1$ to locally Lipschitz. This now allows to analyze the default stepsize change of CMA-ES. We also consider an abstract sampling distribution ν_U which includes multivariate normal distributions as used in CMA-ES.

We summarize the assumptions to prove stability of CMA-ES:

- The objective function is scaling-invariant. This is inherent to our methodology because we define a time-homogeneous Markov chain based upon the normalization of the state variables of CMA-ES.
- The objective function has Lebesgue negligible level sets. This is needed to obtain a lower semicontinuous density for the distribution of the ranked candidate solutions. This is a main assumption to deduce irreducibility from the analysis of an underlying control model.
- The normalization function $R(\cdot)$ is positively homogeneous and continuously Lipschitz, but $R(\cdot)$ may be nonsmooth. This includes natural normalizations, e.g., by the determinant (which is smooth) or an eigenvalue (which is nonsmooth). Positive homogeneity is needed for building the normalized Markov chain and thus (too) inherent to the Markov chain methodology. Lipschitz continuity yields a locally Lipschitz function F for the nonsmooth model (4.1) and allows to connect irreducibility to the analysis of an underlying control model [19].

- The hyperparameter setting assumptions cover all practically relevant algorithm variants (with/without cumulation, with rank-one and rank-mu updates) except when $c_1 + c_\mu = 1$, or when $c_c < 1$ and either $c_\mu = 0$ or $1 - c_c = (1 - c_\sigma)\sqrt{1 - c_1 - c_\mu}$. Without cumulation ($c_c = 1$), the rank-one update is sufficient to prove irreducibility and aperiodicity. However, we need the rank-mu update for our proof when cumulation is used ($c_c < 1$). None of the above cases is important in practice.

Limitations and perspectives We believe that some of the above assumptions can be relaxed with further work, specifically, and based on empirical observations, the assumptions that

- the hyperparameters have to be chosen suitably (in particular $0 < c_\mu < 1$),
- the objective function f has Lebesgue negligible level sets, and
- the sampling distribution is positive and continuous on the entire search space (which is not the case for a distribution on the unit sphere).

In order to conclude—with the approach pursued in this paper—the linear convergence of CMA-ES and its learning of the inverse Hessian, it still remains to be proven that the normalized Markov chain converges geometrically fast to a stationary distribution and satisfies a Law of Large Numbers. This proof could be achieved by finding a potential function for which a geometric drift condition holds [35].

A Proofs in Section 5.2

A.1 Proof of Lemma 5.6

Proof of Lemma 5.6. Let e_k be the k-th vector of the basis $\mathcal{B}$. Let κ be positive and κ' be real. Consider the sequence $\{\theta_t\}_{t=0,1,2,3,4}$ defined by

$$\theta_{t+1} = (z_{t+1}, p_{t+1}, q_{t+1}, \hat{\mathbf{\Sigma}}_{t+1}, r_{t+1}) = F_\Theta\left(\theta_t, \alpha_\Theta\left(\theta_t, v_{t+1}\right)\right)$$

with $v_1 = [\kappa e_k]_{i=1,\dots,\mu}$, $v_2 = -r_1^{-1/2}\Gamma(p_1)^{-1}v_1$, $v_3 = [\kappa' e_k]_{i=1,\dots,\mu}$, and $v_4 = -r_3^{-1/2}\Gamma(p_3)^{-1}v_3$. By Proposition 5.6, we have $v_{1:4} \in \overline{\mathcal{O}^4_{\theta_0}}$ and by Lemma 5.4 we obtain $z_4 = z_2 = 0$. Moreover,

$$q_2 = \kappa \times \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\, r_1^{-1/2}\left[1 - c_c - \Gamma(p_1)^{-1}\right] e_k.$$

Let $\eta \in \mathbb{R}$, and set $\kappa' = \eta \times \left(\kappa\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\, r_1^{-1/2}\left[1 - c_c - \Gamma(p_1)^{-1}\right]\right) = \eta q_2$. Then, similarly to the proof of Lemma 5.5, we have, since $v_3 = \eta[q_2, \dots, q_2]$:

$$q_4 = r_3^{-1/2} \times \left((1-c_c)^2 r_2^{-1/2} + (1 - c_c - \Gamma(p_3)^{-1})\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\,\eta\right) \times q_2 \tag{A.1}$$

where

$$p_3 = (1-c_c)p_2 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\hat{\mathbf{\Sigma}}_2^{-1/2}\mathbf{w}_m^\top v_3$$
$$= (1-c_\sigma)p_2 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\eta\hat{\mathbf{\Sigma}}_2^{-1/2} q_2$$

and thus

$$\Gamma(p_3)^{-1} = \Gamma\left((1-c_\sigma)p_2 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\eta\hat{\mathbf{\Sigma}}_2^{-1/2} q_2\right)^{-1}.$$

We apply the intermediate value theorem to the function

$$\zeta : \eta \mapsto \left[1 - c_c - \Gamma\left((1-c_\sigma)p_2 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\eta\hat{\mathbf{\Sigma}}_2^{-1/2} q_2\right)^{-1}\right] \times \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\,\eta\ ,$$

which is such that $q_4 = r_3^{-1/2} \times \left((1-c_c)^2 r_2^{-1/2} + \zeta(\eta)\right) \times q_2$. Since Γ is continuous by **R**2 and such that when η goes to $\pm\infty$, $\left[1 - c_c - \Gamma\left((1-c_\sigma)p_2 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}} R(\hat{\mathbf{\Sigma}}_2)}\,\eta\hat{\mathbf{\Sigma}}_2^{-1/2} q_2\right)^{-1}\right]$ is strictly positive by **Γ**2, we find

that ζ is continuous and $\zeta(\eta)$ tends to $+\infty$ when η to $+\infty$, and to $-\infty$ when η to $-\infty$. Hence we find $\eta_\kappa \in \mathbb{R}$ (which depends continuously on κ) such that when $\eta = \eta_\kappa$, we have

$$q_4 = 0.$$

For the covariance matrix $\hat{\mathbf{\Sigma}}_1$, we have

$$\hat{\mathbf{\Sigma}}_1 = \frac{(1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu\kappa^2 e_k e_k^\top}{\rho((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu\kappa^2 e_k e_k^\top)} \ ,$$

where $\mathbf{\Sigma}_0 = R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0$. Let $\rho_1 = \rho((1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu\kappa^2 e_k e_k^\top)$ and $\omega_1(\kappa) = c_1c_c(2-c_c)\mu_{\text{eff}}\kappa^2 + c_\mu\kappa^2$ such that

$$\begin{aligned}\hat{\mathbf{\Sigma}}_1 &= \rho_1^{-1}\left[(1-c_1-c_\mu)\mathbf{\Sigma}_0 + c_1c_c(2-c_c)\mu_{\text{eff}}\kappa^2 e_k e_k^\top + c_\mu\kappa^2 e_k e_k^\top\right] \\ &=: \rho_1^{-1}\left[(1-c_1-c_\mu)\mathbf{\Sigma}_0 + \omega_1(\kappa) e_k e_k^\top\right].\end{aligned}$$

The map ω_1 is continuous, with $\omega_1(0) = 0$ and $\omega_1(\kappa) \to \infty$ when $\kappa \to \infty$. Similarly, setting

$$\rho_2 = \rho\left(R(\hat{\mathbf{\Sigma}}_1)^{-1}\hat{\mathbf{\Sigma}}_1 + \left(c_1c_c(2-c_c)\mu_{\text{eff}}\kappa^2\left(1-c_c-\Gamma(p_1)^{-1}\right) + c_\mu\kappa^2\Gamma(p_1)^{-2}\right)e_k e_k^\top\right)$$

we get

$$\begin{aligned}\hat{\mathbf{\Sigma}}_2 &= \rho_2^{-1}\Big[R(\hat{\mathbf{\Sigma}}_1)^{-1}(1-c_1-c_\mu)^2\mathbf{\Sigma}_0 \\ &+ \left(R(\hat{\mathbf{\Sigma}}_1)^{-1}(1-c_1-c_\mu)\omega_1(\kappa) + c_1c_c(2-c_c)\mu_{\text{eff}}\kappa^2\left(1-c_c-\Gamma(p_1)^{-1}\right) + c_\mu\kappa^2\Gamma(p_1)^{-2}\right)e_k e_k^\top\Big] \\ &=: \rho_2^{-1}\left[R(\hat{\mathbf{\Sigma}}_1)^{-1}(1-c_1-c_\mu)^2\mathbf{\Sigma}_0 + \omega_2(\kappa)e_k e_k^\top\right],\end{aligned}$$

with ω_2 continuous since R and Γ are continuous by **R**2 and **Γ**1, $\omega_2(0) = 0$ and $\omega_2(\kappa) \to \infty$ when $\kappa \to \infty$.

Likewise, for the next two steps, we find $\rho_4 > 0$, ω_4 continuous such that $\omega_4(0) = 0$, and $\omega_4(\kappa) \to \infty$ when $\kappa \to \infty$, and

$$\hat{\mathbf{\Sigma}}_4 = \rho_4^{-1}\left[R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4\mathbf{\Sigma}_0 + \omega_4(\kappa)e_k e_k^\top\right].$$

Then, by the intermediate value theorem, there exists $\kappa > 0$ such that

$$\omega_4(\kappa) = R(\hat{\mathbf{\Sigma}}_0)^{-1}R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4(\lambda_{k-1}-\lambda_k) > 0 \ .$$

Therefore,

$$[\hat{\mathbf{\Sigma}}_4]_{\mathcal{B}} = \rho_4^{-1}R(\hat{\mathbf{\Sigma}}_0)^{-1}R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4\text{diag}(\lambda_1,\dots,\lambda_{k-1},\lambda_{k-1},\lambda_{k+1},\dots,\lambda_d) \ .$$

Setting $\gamma = \rho_4^{-1}R(\hat{\mathbf{\Sigma}}_0)^{-1}R(\hat{\mathbf{\Sigma}}_1)^{-1}R(\hat{\mathbf{\Sigma}}_2)^{-1}R(\hat{\mathbf{\Sigma}}_3)^{-1}(1-c_1-c_\mu)^4$, we have proven that we can reach θ_4 with the matrix $\hat{\mathbf{\Sigma}}_4$ defined in (5.19). □

B Proofs in Section 5.3

B.1 Proof of Lemma 5.10

Proof of Lemma 5.10. By **R**3, there exists $\mathbf{C}_0 \in \mathcal{S}_{++}^d$, such that R is differentiable on a neighborhood of $\mathbf{C}_0$. Since R is positively homogeneous by **R**1, then by Lemma 5.8 R is also differentiable on a neighborhood of $\hat{\mathbf{\Sigma}}_0 \coloneqq \rho(\mathbf{C}_0)^{-1}\mathbf{C}_0$. Then, by Corollary 5.1, there exists $\theta_0 = (z_0, p_0, q_0, \hat{\mathbf{\Sigma}}_0, r_0)$ with $z_0 = q_0 = 0$ which is a steadily attracting state.

Let $T \in \mathbb{N}$, $v_{1:T} \in \overline{\mathcal{O}_{\theta_0}^T}$ and $h_{1:T} \in \mathsf{V}^T$. We denote for $t \in \{1,\dots,T\}$:

$$\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) = S_{\theta_0}^t(v_{1:t}) \quad \text{and} \quad \theta_t^h = (z_t^h, p_t^h, q_t^h, \hat{\mathbf{\Sigma}}_t^h, r_t^h) = S_{\theta_0}^t(v_{1:t}+h_{1:t}) \ .$$

We have that, if $v_{1:T} = 0$, then since $q_0 = 0$, $q_t = 0$ for $t = 1,\dots,T$ and using (5.8) we find $\hat{\mathbf{\Sigma}}_t = \hat{\mathbf{\Sigma}}_0$. Since $v \mapsto S_{\theta_0}^t(v)$ is continuous, and since R is differentiable in a neighborhood of $\hat{\mathbf{\Sigma}}_0$, then there exists $M_V > 0$ such that, if $\|v_{1:T}\|^2 \leqslant M_V$, then R is differentiable at $\hat{\mathbf{\Sigma}}_t$. Hence we impose that $\|v_t\|^2 \leqslant M_V/T$ for all $t \in \{1,\dots,T\}$.

Define, for $t = 0, \dots, T$, $b_t = r_0 \times \cdots \times r_t \times R(\hat{\mathbf{\Sigma}}_t)^{-1}$ and $b_t^h = r_0^h \times \dots r_t^h \times R(\hat{\mathbf{\Sigma}}_t^h)^{-1}$, and let $\mathbf{B}_t = b_t \hat{\mathbf{\Sigma}}_t$ and likewise $\mathbf{B}_t^h = b_t^h \hat{\mathbf{\Sigma}}_t^h$. Therefore by positive homogeneity of ρ, $\rho(\mathbf{B}_t) = \rho(b_t \hat{\mathbf{\Sigma}}_t) = b_t \rho(\hat{\mathbf{\Sigma}}_t) = b_t$ since $\rho(\hat{\mathbf{\Sigma}}_t) = 1$. Similarly $\rho(\mathbf{B}_t^h) = b_t^h$ and thus

$$\hat{\mathbf{\Sigma}}_t = \frac{\mathbf{B}_t}{\rho(\mathbf{B}_t)} \quad \text{and} \quad \hat{\mathbf{\Sigma}}_t^h = \frac{\mathbf{B}_t^h}{\rho(\mathbf{B}_t^h)} \ . \tag{B.1}$$

Moreover, define $\tilde{q}_t = \sqrt{\tilde{r}_{t-1}} q_t$ and $\tilde{q}_t^h = \sqrt{\tilde{r}_{t-1}^h} q_t^h$ as well as $\tilde{v}_{t+1} = \sqrt{\tilde{r}_t} v_{t+1}$ and $\tilde{h}_{t+1} = \sqrt{\tilde{r}_t^h} h_{t+1}$ where $\tilde{r}_t = r_0 \times \cdots \times r_t$ and $\tilde{r}_t^h = r_0^h \times \cdots \times r_t^h$. Hence, by applying (5.7):

$$\begin{aligned}\tilde{q}_{t+1} &= \sqrt{\tilde{r}_t} r_t^{-1/2}(1-c_c) q_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \sqrt{\tilde{r}_t} \mathbf{w}_m^\top v_{t+1} \\ &= (1-c_c)\sqrt{\tilde{r}_{t-1}} q_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \mathbf{w}_m^\top \tilde{v}_{t+1} = (1-c_c)\tilde{q}_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \mathbf{w}_m^\top \tilde{v}_{t+1}\end{aligned}$$

and likewise

$$\tilde{q}_{t+1}^h = (1-c_c)\tilde{q}_t^h + \sqrt{c_c(2-c_c)\mu_{\text{eff}}} \mathbf{w}_m^\top \left((\tilde{r}_t^h / \tilde{r}_t)^{1/2} \tilde{v}_{t+1} + \tilde{h}_{t+1} \right) \ . \tag{B.2}$$

Denote $\mathbf{A}_{t+1}$ such that $\hat{\mathbf{\Sigma}}_{t+1} = \mathbf{A}_{t+1}/\rho(\mathbf{A}_{t+1})$ in (5.8). (Alternatively the matrix $\tilde{\mathbf{\Sigma}}_{t+1}$ in (2.16) equals $\mathbf{A}_{t+1}$). Then by positive homogeneity of R, $R(\hat{\mathbf{\Sigma}}_{t+1}) = R(\mathbf{A}_{t+1})/\rho(\mathbf{A}_{t+1})$ such that

$$\frac{\mathbf{A}_{t+1}}{R(\mathbf{A}_{t+1})} = \frac{\mathbf{A}_{t+1}}{\rho(\mathbf{A}_{t+1}) R(\hat{\mathbf{\Sigma}}_{t+1})} = \frac{\hat{\mathbf{\Sigma}}_{t+1}}{R(\hat{\mathbf{\Sigma}}_{t+1})} \tag{B.3}$$

Then, using the previous equation and (5.8):

$$\begin{aligned}\mathbf{B}_{t+1} &= b_{t+1}\hat{\mathbf{\Sigma}}_{t+1} = r_0 \times \cdots \times r_{t+1} \times R(\hat{\mathbf{\Sigma}}_{t+1})^{-1} \hat{\mathbf{\Sigma}}_{t+1} = \tilde{r}_{t+1} R(\hat{\mathbf{\Sigma}}_{t+1})^{-1} \hat{\mathbf{\Sigma}}_{t+1} && \text{(B.4)} \\ &= \tilde{r}_{t+1} R(\mathbf{A}_{t+1})^{-1} \mathbf{A}_{t+1} && \text{(B.5)} \\ &= \underbrace{\tilde{r}_{t+1} \times r_{t+1}^{-1}}_{\tilde{r}_t} \times \left((1-c_1-c_\mu) R(\hat{\mathbf{\Sigma}}_t)^{-1} \hat{\mathbf{\Sigma}}_t + c_1 q_{t+1} q_{t+1}^\top + c_\mu \sum_{i=1}^{\mu} w_i^c v_{t+1}^i (v_{t+1}^i)^\top \right) && \text{(B.6)} \\ &= (1-c_1-c_\mu)\tilde{r}_t R(\hat{\mathbf{\Sigma}}_t)^{-1} \hat{\mathbf{\Sigma}}_t + c_1 \tilde{r}_t q_{t+1} q_{t+1}^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \tilde{r}_t v_{t+1}^i (v_{t+1}^i)^\top && \text{(B.7)} \\ &= (1-c_1-c_\mu)\mathbf{B}_t + c_1 \tilde{q}_{t+1} \tilde{q}_{t+1}^\top + c_\mu \sum_{i=1}^{\mu} w_i^c \tilde{v}_{t+1}^i (\tilde{v}_{t+1}^i)^\top \ . && \text{(B.8)}\end{aligned}$$

Likewise,

$$\mathbf{B}_{t+1}^h = (1-c_1-c_\mu)\mathbf{B}_t^h + c_1 \tilde{q}_{t+1}^h (\tilde{q}_{t+1}^h)^\top + c_\mu \tilde{r}_t^h \sum_{i=1}^{\mu} w_i^c (v_{t+1}^i + h_{t+1}^i)(v_{t+1}^i + h_{t+1}^i)^\top \ . \tag{B.9}$$

Let $s = d(d+1)/2$ be the dimension of the set of symmetric matrices $\mathcal{S}^d$ as a real vector space. Let $\psi_1 \in \mathbb{R}^d$ be a nonzero vector and define then $\psi_2, \dots, \psi_s$ nonzero vectors of $\mathbb{R}^d$, such that $(\psi_1\psi_1^\top, \dots, \psi_s\psi_s^\top)$ forms a basis of $\mathcal{S}^d$. Scaling down the length of ψ_k does not change that we have a basis of $\mathcal{S}^d$ and thus we impose that $\|\psi_k\| \leqslant \varepsilon$, where ε is a positive constant that we precise in the next paragraph. Set $T = 2s(s-1) + 4$ and set $v_{1:T}$ as below.

For $t \in \{0, \dots, s-1\}$, we set

$$v_{2t+1} = \tilde{r}_{2t}^{-1/2}[\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu} \text{ and } v_{2t+2} = -(1-c_c)\tilde{r}_{2t+1}^{-1/2}[\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu} \tag{B.10}$$

such that

$$\tilde{v}_{2t+1} = [\psi_{t+1}, \dots, \psi_{t+1}] \text{ and } \tilde{v}_{2t+2} = -(1-c_c)[\psi_{t+1}, \dots, \psi_{t+1}]. \tag{B.11}$$

Moreover, we choose $\varepsilon > 0$ small enough so that $\|v_k\|^2 \leqslant M_V/T$ for all $k = 1, \dots, 2s$. By definition of M_V earlier in the proof, we have that R is differentiable in $\hat{\mathbf{\Sigma}}_t$ for $t = 0, \dots, T$. If moreover $p_t \neq 0$ for $t = 1, \dots, T$, then by Lemma 5.9, $v \to S_{\theta_0}^T(v)$ is differentiable in $v_{1:T}$. Besides, by Proposition 5.6, we have $v_{1:2s} \in \overline{\mathcal{O}_{\theta_0}^{2s}}$.

Observe now that there exists $\psi_1 \in \mathbb{R}^d$ such that $\|\psi_1\| \leqslant \varepsilon$ and p_1, p_2 are nonzero. Indeed, $p_1 = (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\mathbf{\Sigma}}_0)^{1/2}\hat{\mathbf{\Sigma}}_0^{-1/2}\sum_{i=1}^{\mu} w_i^m v_1^i$ and using $\mathbf{B}_0 = r_0 R(\hat{\mathbf{\Sigma}}_0)^{-1}\hat{\mathbf{\Sigma}}_0$ we find $p_1 = (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}r_0^{1/2}\mathbf{B}_0^{-1/2}\sum_{i=1}^{\mu} w_i^m \tilde{r}_0^{-1/2}\psi_1 = (1-c_\sigma)p_0 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{B}_0^{-1/2}\psi_1$ and

$$p_2 = (1-c_\sigma)p_1 + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}R(\hat{\mathbf{\Sigma}}_1)^{1/2}\hat{\mathbf{\Sigma}}_1^{-1/2}\sum_{i=1}^{\mu} w_i^m v_2^i \tag{B.12}$$

$$= (1-c_\sigma)^2 p_0 + (1-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{B}_0^{-1/2}\psi_1 \tag{B.13}$$

$$- \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\tilde{r}_1^{1/2}\mathbf{B}_1^{-1/2}\sum w_i^m(1-c_c)\tilde{r}_1^{-1/2}\psi_1 \tag{B.14}$$

$$= (1-c_\sigma)^2 p_0 + [(1-c_\sigma)\mathbf{B}_0^{-1/2} - (1-c_c)\mathbf{B}_1^{-1/2}]\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\psi_1 \ . \tag{B.15}$$

Since $v_2 = -(1-c_c)\tilde{r}_1^{-1/2}[\psi_1, \dots, \psi_1]$ with $\psi_1 \neq 0$, given that according to (B.8), $\mathbf{B}_1 = \alpha_1\mathbf{B}_0 + \alpha_2\psi_1\psi_1^\top + \alpha_3 q_1 q_1^\top$, for $\alpha_1, \alpha_2, \alpha_3$ some nonnegative constants and $\alpha_2 + \alpha_3 > 0$ since $c_1 + c_\mu > 0$, and $\psi_1, q_1 \neq 0$ (see below), we have $(1-c_\sigma)\mathbf{B}_0^{-1/2} \neq (1-c_c)\mathbf{B}_1^{-1/2}$. Moreover, up to scaling $\psi_2, \dots, \psi_s$ sufficiently smaller than ψ_1, we can ensure that $p_t \neq 0$ for $t \in \{3, \dots, 2s\}$. Then, by Lemma 5.9, and by composition since $S_{\theta_0}^{t+1}(v_{1:t+1}) = S_{S_{\theta_0}^t(v_{1:t})}^1(v_{t+1})$ we find by induction that $S_{\theta_0}^{2s}$ is differentiable at $v_{1:2s}$. Then, by induction, since $\tilde{q}_{t+1} = (1-c_c)\tilde{q}_t + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top \tilde{v}_{t+1}$ with $\tilde{v}_{2t+1} = [\psi_{t+1}, \dots, \psi_{t+1}]$ and $\tilde{v}_{2t+2} = -(1-c_c)[\psi_{t+1}, \dots, \psi_{t+1}]$, we find that, for every $t \in \{0, \dots, s-1\}$, we have:

$$\tilde{q}_{2t+1} = \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\psi_{t+1} \text{ and } q_{2t+2} = 0. \tag{B.16}$$

For $t = 0, \dots, s-1$, let $\kappa_t^1 \in \mathbb{R}$ be arbitrary (we fix the value of κ_t^1 later in the proof). We set, given an arbitrary real number $\varepsilon_1 \in \mathbb{R}$, for $t = 0, \dots, s-1$:

$$h_{2t+1} = [(\tilde{r}_{2t}^h)^{-1/2} - \tilde{r}_{2t}^{-1/2} + (\tilde{r}_{2t}^h)^{-1/2}\kappa_t^1\varepsilon_1] \times [\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu}$$

which implies

$$\tilde{h}_{2t+1} = [1 - (\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \kappa_t^1\varepsilon_1] \times [\psi_{t+1}, \dots, \psi_{t+1}] \tag{B.17}$$

and

$$h_{2t+2} = -(1-c_c)[(\tilde{r}_{2t+1}^h)^{-1/2} - \tilde{r}_{2t+1}^{-1/2} + (\tilde{r}_{2t+1}^h)^{-1/2}\kappa_t^1\varepsilon_1] \times [\psi_{t+1}, \dots, \psi_{t+1}] \in \mathbb{R}^{d\mu} \ ,$$

so that, by induction, starting from (B.2) we have:

$$\tilde{q}_{2t+1}^h = (1-c_c)\tilde{q}_{2t}^h + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top((\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2}\tilde{v}_{2t+1} + \tilde{h}_{2t+1}) \tag{B.18}$$

$$= (1-c_c) \times 0 \tag{B.19}$$

$$+ \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top\left((\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \left(1 - (\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \kappa_t^1\varepsilon_1\right)\right)[\psi_{t+1}, \dots, \psi_{t+1}] \tag{B.20}$$

$$= \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\left(1 + \kappa_t^1\varepsilon_1\right)\psi_{t+1} \tag{B.21}$$

and

$$\begin{aligned}
\tilde{q}_{2t+2}^h &= (1-c_c)\tilde{q}_{2t+1}^h + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top((\tilde{r}_{2t+1}^h/\tilde{r}_{2t+1})^{1/2}\tilde{v}_{2t+2} + \tilde{h}_{2t+2}) \\
&= (1-c_c)\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\left(1 + \kappa_t^1\varepsilon_1\right)\psi_{t+1} \\
&\quad - (1-c_c)\sqrt{c_c(2-c_c)\mu_{\text{eff}}}\mathbf{w}_m^\top\left(\sqrt{\frac{\tilde{r}_{2t+1}^h}{\tilde{r}_{2t+1}}} + \left(1 - \sqrt{\frac{\tilde{r}_{2t+1}^h}{\tilde{r}_{2t+1}}} + \kappa_t^1\varepsilon_1\right)\right)[\psi_{t+1}, \dots, \psi_{t+1}] \\
&= 0
\end{aligned}$$

Note that, for $i = 1, \dots, \mu$:

$$(\tilde{r}_{2t}^h)^{1/2} \times (v_{2t+1}^i + h_{2t+1}^i) = \left((\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \left(1 - (\tilde{r}_{2t}^h/\tilde{r}_{2t})^{1/2} + \kappa_t^1\varepsilon_1\right)\right)\psi_{t+1} = \left(1 + \kappa_t^1\varepsilon_1\right)\psi_{t+1} \ .$$

Then, using (B.9), we obtain for $t \in \{0, \dots, s-1\}$, when $\varepsilon_1 \to 0$:

$$\mathbf{B}_{2t+1}^h = (1-c_1-c_\mu)\mathbf{B}_{2t}^h + c_1c_c(2-c_c)\mu_{\text{eff}}(1+\kappa_t^1\varepsilon_1)^2\psi_{t+1}\psi_{t+1}^\top \tag{B.22}$$

$$+ c_\mu\sum_{i=1}^{\mu} w_i^c\left(1 + \kappa_t^1\varepsilon_1\right)^2\psi_{t+1}\psi_{t+1}^\top \tag{B.23}$$

$$= (1-c_1-c_\mu)\mathbf{B}_{2t}^h + [c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu] \times \left(1 + 2\kappa_t^1\varepsilon_1\right)\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \tag{B.24}$$

From (B.8), we have that

$$\begin{aligned}\mathbf{B}_{2t+1} &= (1-c_1-c_\mu)\mathbf{B}_{2t} + c_1\tilde{q}_{2t+1}\tilde{q}_{2t+1}^\top + c_\mu\sum_{i=1}^{\mu} w_i^c \tilde{v}_{2t+1}^i(\tilde{v}_{2t+1}^i)^\top \\ &= (1-c_1-c_\mu)\mathbf{B}_{2t} + c_1c_c(2-c_c)\mu_{\text{eff}}\psi_{t+1}\psi_{t+1}^\top + c_\mu\sum_{i=1}^{\mu} w_i^c\psi_{t+1}\psi_{t+1}^\top \\ &= (1-c_1-c_\mu)\mathbf{B}_{2t} + \left[c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu\right]\psi_{t+1}\psi_{t+1}^\top\end{aligned}$$

that we use in (B.24) to obtain

$$\begin{aligned}\mathbf{B}_{2t+1}^h &= \mathbf{B}_{2t+1} + (1-c_1-c_\mu)\left(\mathbf{B}_{2t}^h - \mathbf{B}_{2t}\right) + \underbrace{\left[c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu\right]\times 2}_{:=c_b}\kappa_t^1\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \\ &= \mathbf{B}_{2t+1} + (1-c_1-c_\mu)\left(\mathbf{B}_{2t}^h - \mathbf{B}_{2t}\right) + c_b\kappa_t^1\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1)\ . \end{aligned} \tag{B.25}$$

Moreover, for $i = 1, \ldots, \mu$:

$$\begin{aligned}(\tilde{r}_{2t+1}^h)^{1/2}\times(v_{2t+2}^i + h_{2t+2}^i) &= -(1-c_c)\left((\tilde{r}_{2t+1}^h/\tilde{r}_{2t+1})^{1/2} + \left(1-(\tilde{r}_{2t+1}^h/\tilde{r}_{2t+1})^{1/2} + \kappa_t^1\varepsilon_1\right)\right)\psi_{t+1} \\ &= -(1-c_c)\left(1+\kappa_t^1\varepsilon_1\right)\psi_{t+1}\ .\end{aligned}$$

Thus, we obtain, by (B.9) and (B.25):

$$\begin{aligned}\mathbf{B}_{2t+2}^h &= (1-c_1-c_\mu)\mathbf{B}_{2t+1}^h + c_\mu\sum_{i=1}^{\mu} w_i^c(1-c_c)^2(1+\kappa_t^1\varepsilon_1)^2\psi_{t+1}\psi_{t+1}^\top \\ &= (1-c_1-c_\mu)\mathbf{B}_{2t+1} + (1-c_1-c_\mu)^2\left(\mathbf{B}_{2t}^h - \mathbf{B}_{2t}\right) + (1-c_1-c_\mu)c_b\kappa_t^1\varepsilon_1\psi_{t+1}\psi_{t+1}^\top \\ &\quad + c_\mu(1-c_c)^2(1+\kappa_t^1\varepsilon_1)^2\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1) \\ &= (1-c_1-c_\mu)\mathbf{B}_{2t+1} + c_\mu(1-c_c)^2\psi_{t+1}\psi_{t+1}^\top + (1-c_1-c_\mu)^2\left(\mathbf{B}_{2t}^h - \mathbf{B}_{2t}\right) \\ &\quad + d_b\kappa_t^1\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1)\end{aligned}$$

with

$$d_b = (1-c_1-c_\mu)c_b + 2c_\mu(1-c_c)^2 = 2(1-c_1-c_\mu)\left(c_1c_c(2-c_c)\mu_{\text{eff}} + c_\mu\right) + 2c_\mu(1-c_c)^2\ .$$

Yet, by (B.8) since by (B.16) $q_{2t+2} = 0$ and by (B.11) $\tilde{v}_{2t+2}^i = -(1-c_c)\psi_{t+1}$:

$$\mathbf{B}_{2t+2} = (1-c_1-c_\mu)\mathbf{B}_{2t+1} + c_\mu(1-c_c)^2\psi_{t+1}\psi_{t+1}^\top\ .$$

Therefore,

$$\mathbf{B}_{2t+2}^h - \mathbf{B}_{2t+2} = (1-c_1-c_\mu)^2\left(\mathbf{B}_{2t}^h - \mathbf{B}_{2t}\right) + d_b\kappa_t^1\varepsilon_1\psi_{t+1}\psi_{t+1}^\top + o(\varepsilon_1)\ .$$

Then, by induction, we get,

$$\mathbf{B}_{2s}^h = \mathbf{B}_{2s} + \varepsilon_1\sum_{t=1}^{s}(1-c_1-c_\mu)^{2s-2t}d_b\kappa_{t-1}^1\psi_t\psi_t^\top + (1-c_1-c_\mu)^{2s}(\mathbf{B}_0^h - \mathbf{B}_0) + o(\varepsilon_1)\ ,$$

with $\mathbf{B}_0^h - \mathbf{B}_0 = 0$ by definition. By induction on $k \in \{1, \ldots, s-2\}$, we set for $t \in \{0, \ldots, s-1\}$:

$$v_{2t+2ks+1} = \tilde{r}_{2t+2ks}^{-1/2}[\psi_{t+1}, \ldots, \psi_{t+1}] \in \mathbb{R}^{d\mu}\ ,$$

and

$$v_{2t+2ks+2} = -(1-c_c)\tilde{r}_{2t+2ks+1}^{-1/2}[\psi_{t+1}, \ldots, \psi_{t+1}] \in \mathbb{R}^{d\mu}\ .$$

We also set, given arbitrary real numbers $\varepsilon_k \in \mathbb{R}$ and for some $\kappa_t^k \in \mathbb{R}$ for $t \in \{0, \ldots, s-1\}$:

$$h_{2t+2ks+1} = [(\tilde{r}_{2t+2ks}^h)^{-1/2} - \tilde{r}_{2t+2ks}^{-1/2} + (\tilde{r}_{2t+2ks}^h)^{-1/2}\kappa_t^s\varepsilon_1]\times[\psi_{t+1}, \ldots, \psi_{t+1}] \in \mathbb{R}^{d\mu}$$

and

$$h_{2t+2ks+2} = -(1-c_c)[(\tilde{r}_{2t+2ks+1}^h)^{-1/2} - \tilde{r}_{2t+2ks+1}^{-1/2} + (\tilde{r}_{2t+2ks+1}^h)^{-1/2}\kappa_t^s\varepsilon_1]\times[\psi_{t+1}, \ldots, \psi_{t+1}] \in \mathbb{R}^{d\mu}\ .$$

Then, similarly to above, we obtain $q_{2(k+1)s} = 0$ and

$$\mathbf{B}^h_{2(k+1)s} = \mathbf{B}_{2(k+1)s} + \varepsilon_k \sum_{t=1}^{s} (1-c_1-c_\mu)^{2s-2t} d_b \kappa_{t-1}^k \psi_t \psi_t^\top + (1-c_1-c_\mu)^{2s}(\mathbf{B}^h_{2ks} - \mathbf{B}_{2ks}) + o(\varepsilon_k) \ .$$

Thus, by induction, we get $q^h_{2s(s-1)} = 0$ and

$$\mathbf{B}^h_{2s(s-1)} = \mathbf{B}_{2s(s-1)} + \sum_{k=1}^{s-1} \varepsilon_k (1-c_1-c_\mu)^{2s(s-1)-2ks} \sum_{t=1}^{s} (1-c_1-c_\mu)^{2s-2t} d_b \kappa_{t-1}^k \psi_t \psi_t^\top + o(\varepsilon_{1:s-1})$$

Note moreover that we can assume again that $p_t \neq 0$, up to choosing again the ψ_k, $k \geqslant 2$, sufficiently smaller than ψ_1.

By Lemmas 5.3 and 5.5, for any $v_{2s(s-1)+1} \in \overline{\mathcal{O}^1_{\theta_{2s(s-1)}}}$, there exists $v_{2s(s-1)+2:2s(s-1)+4} \in \overline{\mathcal{O}^3_{\theta_{2s(s-1)+1}}}$ such that $z_{2s(s-1)+4} = q_{2s(s-1)+4} = 0$. Moreover, when $v_{2s(s-1)+1} \to 0$, then we have that $z_{2s(s-1)+1}$ and $q_{2s(s-1)+1}$ tend to 0 and thus we can impose that $v_{2s(s-1)+2:2s(s-1)+4} \to 0$ as well. In particular, we can choose $v_{2s(s-1)+1}$ small enough such that $p_t \neq 0$ for $t = 2s(s-1)+1, \dots, 2s(s-1)+4$. Hence, by Lemma 5.9, $S_{\theta_0}^{2s(s-1)+4}$ is differentiable at $v_{2s(s-1)+4}$ (we have that R is differentiable at $\hat{\mathbf{\Sigma}}_t$ for all $t = 1, \dots, T$ by imposing $v_{1:T}$ small enough, see the beginning of the proof).

Consider then $(\mathbf{S}_1, \dots, \mathbf{S}_{s-1})$ a basis of $\ker \mathcal{D}\rho(\mathbf{B}_{2s(s-1)+4})$. For $k = 1, \dots, s-1$, we can choose then the $\kappa_t^k \in \mathbb{R}$, $t = 0, \dots, s-1$ so that we have

$$(1-c_1-c_\mu)^{2s(s-1)-2ks} \sum_{t=1}^{s} (1-c_1-c_\mu)^{2s-2t} d_b \kappa_{t-1}^k \psi_t \psi_t^\top = \mathbf{S}_k \ . \tag{B.26}$$

This is possible since $(\psi_1\psi_1^\top, \dots, \psi_s\psi_s^\top)$ is a basis of $\mathcal{S}^d$, and by the intermediate value theorem applied to the LHS of (B.26), for $k = 1, \dots, s$. Set $T = 2s(s-1)+4$. Then, we have, when $\varepsilon_{1:s-1} \to 0$,

$$\mathbf{B}^h_T = \mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}) \ . \tag{B.27}$$

Therefore, since $\mathbf{S}_k \in \ker \mathcal{D}\rho(\mathbf{B}_T)$ for $k = 1, \dots, s$, we have

$$\rho\left(\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1})\right) = \rho(\mathbf{B}_T) + \sum\nolimits_{k=1}^{s-1} \varepsilon_k \underbrace{\mathcal{D}\rho(\mathbf{B}_T)\mathbf{S}_k}_{=0} + o(\varepsilon_{1:s-1}) = \rho(\mathbf{B}_T) + o(\varepsilon_{1:s-1})$$

and using (B.1) and (B.27)

$$\begin{aligned}
\hat{\mathbf{\Sigma}}^h_T &= \frac{\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1})}{\rho\left(\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1})\right)} = \frac{\mathbf{B}_T + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k}{\rho\left(\mathbf{B}_T\right)} + o(\varepsilon_{1:s-1}) \\
&= \hat{\mathbf{\Sigma}}_T + \sum_{k=1}^{s-1} \varepsilon_k \rho\left(\mathbf{B}_T\right)^{-1} \mathbf{S}_k + o(\varepsilon_{1:s-1}).
\end{aligned}$$

However, $(\mathbf{S}_1, \dots, \mathbf{S}_{s-1})$ is a basis of $\ker \mathcal{D}\rho(\mathbf{B}_T)$, and by Lemma 5.8, $\ker \mathcal{D}\rho(\mathbf{B}_T) = \ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_T) = \mathrm{T}_{\hat{\mathbf{\Sigma}}_T} \rho^{-1}(\{1\})$. Thus we have shown that for every $h_{\mathbf{\Sigma}} \in T_{\hat{\mathbf{\Sigma}}_T} \rho^{-1}(\{1\})$ for which we can find ε_k such that $h_{\mathbf{\Sigma}} = \sum_{k=1}^{s-1} \varepsilon_k \rho\left(\mathbf{B}_T\right)^{-1} \mathbf{S}_k$, there exist $h_z, h_p \in \mathbb{R}^d$, $h_r \in \mathbb{R}$ and $h_{1:T} \in (\mathbb{R}^{d\mu})^T$ such that $\mathcal{D}S^T_{\theta_0}(v_{1:T})h_{1:T} = [h_z, h_p, 0, h_{\mathbf{\Sigma}}, h_t]$ which is the statement (ii) of the lemma. The statements (i) and (iii) have been proven earlier in the proof. □

B.2 Proof of Lemma 5.11

Proof of Lemma 5.11. Let $\theta_0 \in \mathsf{X}$ be a steadily attracting state satisfying Lemma 5.10. Then, there exists $T_0 > 0$ and $v_{1:T0} \in \overline{\mathcal{O}^T_{\theta_0}}$ such that conditions (i), (ii), (iii) of Lemma 5.10 are satisfied. Let $T > T_0$, $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$ and $h_{1:T} \in \mathsf{V}^T$. We denote for every $t \in \{1, \dots, T\}$

$$\theta_t = (z_t, p_t, q_t, \hat{\mathbf{\Sigma}}_t, r_t) \coloneqq S^t_{\theta_0}(v_{1:t}) \quad \text{and} \quad \theta^h_t = (z^h_t, p^h_t, q^h_t, \hat{\mathbf{\Sigma}}^h_t, r^h_t) \coloneqq S^t_{\theta_0}(v_{1:t} + h_{1:t}).$$

Let $s = d(d+1)/2$ be the dimension of $\mathcal{S}^d$. Then, $\ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_{T_0}) = \mathrm{T}_{\hat{\mathbf{\Sigma}}_{T_0}} \rho^{-1}(\{1\})$ is a vector space of dimension $s-1$. Let $(\mathbf{S}_1, \dots, \mathbf{S}_{s-1})$ be a basis of $\ker \mathcal{D}\rho(\hat{\mathbf{\Sigma}}_{T_0})$. Then for $k = 1, \dots, s-1$, by condition (ii) in Lemma 5.10, there exists $\xi^k_{1:T_0} \in \mathsf{V}^{T_0}$ such that $\mathcal{D}S^{T_0}_{\theta_0}(v_{1:T_0})\xi^k_{1:T_0} = [h_z, h_p, 0, \mathbf{S}_k, h_r]$ (for some h_z, h_p, h_r). If $h_{1:T_0} = \varepsilon_k \xi^k_{1:T_0} \in \mathsf{V}^{T_0}$ for $\varepsilon_k \in \mathbb{R}$, we have by Taylor expansion and linearity of the differential:

$$\hat{\mathbf{\Sigma}}^h_{T_0} = \hat{\mathbf{\Sigma}}_{T_0} + \varepsilon_k \mathbf{S}_k + o(\varepsilon_k) \ .$$

Set then $h_{1:T_0} = \sum_{k=1}^{s-1} \varepsilon_k \xi^k_{1:T_0}$, so that, by linearity of the differential:

$$\hat{\mathbf{\Sigma}}^h_{T_0} = \hat{\mathbf{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}) \ . \tag{B.28}$$

Moreover, by conditions (ii) and (iii) in Lemma 5.10, we get

$$z_{T_0} = q_{T_0} = 0 \quad \text{and} \quad p \coloneqq p_{T_0} \neq 0 \ , \tag{B.29}$$

and by Taylor expansion since $S^{T_0}_{\theta_0}(v_{1:T_0} + h_{1:T_0}) = S^{T_0}_{\theta_0}(v_{1:T_0}) + \sum_{k=1}^{s-1} \varepsilon_k \mathcal{D}S^{T_0}_{\theta_0}(v_{1:T_0})\xi^k_{1:T_0} + o(\varepsilon_{1:s-1})$

$$z^h_{T_0} = 0 + O(\varepsilon_{1:s-1}) \quad \text{and} \quad q^h_{T_0} = 0 + o(\varepsilon_{1:s-1}) \quad \text{and} \quad p^h_{T_0} = p + O(\varepsilon_{1:s-1}) \ . \tag{B.30}$$

Let $j \in \mathbb{N}$ and set $T = T_0 + j + 5$ and

$$v_{T_0+1:T_0+j+5} = 0 \tag{B.31}$$

Then $v_{T_0+1:T_0+j+5} = 0 \in \overline{\mathcal{O}^{j+5}_{\theta_{T_0}}}$ by Proposition 5.6. Since $z_{T_0} = q_{T_0} = 0$, then we obtain by applying the update equations (5.5) and (5.7), that $z_{T_0:T} = q_{T_0:T} = 0$. By condition (i) in Lemma 5.10, $S^{T_0}_{\theta_0}$ is differentiable at $v_{1:T_0}$. Moreover, for $t = T_0, \dots, T-1$, we have $z_t = q_t = 0$ and $v_{t+1} = 0$, hence by Lemma 5.9 (case a) $S^1_{\theta_t}$ is differentiable at v_{t+1}. By chain rule, $S^T_{\theta_0}$ is differentiable at $v_{1:T}$.

We set $h_{T_0+1:T_0+j} = 0$, then since $\theta^h_{T_0+j} = S^j_{\theta^h_{T_0}}(v_{T_0+1:T_0+j})$ by applying (5.5), (5.6) and (5.7) with $v = 0$ and using (B.30), we have

$$z^h_{T_0+j} = 0 + O(\varepsilon_{1:s-1}) \quad \text{and} \quad q^h_{T_0+j} = 0 + o(\varepsilon_{1:s-1}) \quad \text{and} \quad p^h_{T_0+j} = (1-c_\sigma)^j p + O(\varepsilon_{1:s-1}) \ . \tag{B.32}$$

Moreover, set

$$\begin{cases} h_{T_0+j+1} = (H_1, \dots, H_1) \in \mathbb{R}^{d\mu} & \text{for some } H_1 \in \mathbb{R}^d \\ h_{T_0+j+2} = -(1-c_1-c_\mu)^{-1/2}\Gamma(p_{T_0+j+1})^{-1} h_{T_0+j+1} & \\ h_{T_0+j+3} = (H_3, \dots, H_3) \in \mathbb{R}^{d\mu} & \text{for some } H_3 \in \mathbb{R}^d \\ h_{T_0+j+4} = -(1-c_1-c_\mu)^{-1/2}\Gamma(p_{T_0+j+3})^{-1} h_{T_0+j+3} & \\ h_{T_0+j+5} = (H_5, \dots, H_5) \in \mathbb{R}^{d\mu} & \text{for some } H_5 \in \mathbb{R}^d \end{cases}$$

Note that $p_{T_0+k} = (1-c_\sigma)^k p$ for $k = 1, \dots, j+5$ by (5.6), and by Taylor expansion:

$$p^h_{T_0+j+1} = (1-c_\sigma)^{j+1} p + O(\varepsilon_{1:s-1}, H_1) \ .$$

Yet, Γ is locally Lipschitz by $\mathbf{\Gamma}1$, thus $\Gamma(p^h_{T_0+j+1}) = \Gamma((1-c_\sigma)^{j+1}p) + O(\varepsilon_{1:s-1}, H_1)$. Moreover, since by $\mathbf{R}1$, we have $r_{T_0+k} = R((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_{T_0+k})^{-1}\hat{\mathbf{\Sigma}}_{T_0+k}) = 1-c_1-c_\mu$ and since R is locally Lipschitz by $\mathbf{R}2$, we have

$$\begin{aligned} r^h_{T_0+k} = R((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}_{T_0+k})^{-1}\hat{\mathbf{\Sigma}}_{T_0+k} + O(h_{1:T_0+k})) &= 1-c_1-c_\mu + O(h_{1:T_0+k}) \\ &= 1-c_1-c_\mu + O(\varepsilon_{1:s-1}) + O(h_{T_0+1:k}) \ . \end{aligned} \tag{B.33}$$

When $H_1, H_3, H_5, \varepsilon_{1:s-1} \to 0$, since

$$\theta^h_{T_0+j+1} = S^1_{\theta^h_{T_0+j}}(0 + h_{T_0+j+1}) \tag{B.34}$$

by applying (5.5) we find

$$z^h_{T_0+j+1} = \frac{z^h_{T_0+j} + c_m H_1}{\sqrt{r^h_{T_0+j+1}}\Gamma(p^h_{T_0+j+1})}$$

and thus using (B.32) and (B.33):

$$z^h_{T_0+j+1} = c_m(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1 + O(\varepsilon_{1:s-1}) + o(H_1)$$

so,

$$\begin{aligned} z^h_{T_0+j+2} &= \frac{z^h_{T_0+j+1} - c_m(1-c_1-c_\mu)^{-1/2}\Gamma(p_{T_0+j+1})^{-1}H_1}{\sqrt{r^h_{T_0+j+2}}\Gamma(p^h_{T_0+j+2})} \\ &= O(\varepsilon_{1:s-1}) + o(H_1) \ . \end{aligned}$$

Likewise,

$$z^h_{T_0+j+4} = O(\varepsilon_{1:s-1}) + o(H_1, H_3),$$

so that, in the end, since R is locally Lipschitz by **R**2 and Γ is locally Lipschitz by **Γ**1, then

$$z^h_T = z^h_{T_0+j+5} = O(\varepsilon_{1:s-1}) + o(H_1, H_3) + c_m r_T^{-1/2}\Gamma(p_T)^{-1}H_5 + o(H_5).$$

Furthermore using (B.34) and (B.32),

$$\begin{aligned} q^h_{T_0+j+1} &= (1-c_c)(r^h_{T_0+j})^{-1/2}q^h_{T_0+j} + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(0+H_1) \\ &= \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_1 + o(\varepsilon_{1:s-1}) \end{aligned}$$

and

$$\begin{aligned} q^h_{T_0+j+2} &= (1-c_c)(r^h_{T_0+j+1})^{-1/2}q^h_{T_0+j+1} \\ &\quad + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(0 - (1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1) \\ &= (1-c_c)(1-c_1-c_\mu+O(\varepsilon_{1:s-1},H_1))^{-1/2}\sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_1 + o(\varepsilon_{1:s-1}) \\ &\quad + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1 \\ &= \underbrace{(1-c_1-c_\mu)^{-1/2}\left[1-c_c-\Gamma((1-c_\sigma)^{j+1}p)^{-1}\right]\sqrt{c_c(2-c_c)\mu_{\text{eff}}}}_{=:d^p_{j+1}}H_1 + o(\varepsilon_{1:s-1},H_1) \ . \end{aligned}$$

Likewise,

$$\begin{aligned} q^h_{T_0+j+3} &= (r_{T_0+j+2})^{-1/2}(1-c_c)q^h_{T_0+j+2} + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_3 \\ &= (1-c_1-c_\mu)^{-1/2}(1-c_c)d^p_{j+1}H_1 + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_3 + o(\varepsilon_{1:s-1},H_1) \end{aligned}$$

and

$$\begin{aligned} q^h_{T_0+j+4} &= (r_{T_0+j+3})^{-1/2}(1-c_c)q^h_{T_0+j+3} - \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3})^{-1}H_3 \\ &= (1-c_1-c_\mu)^{-1}(1-c_c)^2d^p_{j+1}H_1 + (1-c_1-c_\mu)^{-1/2}(1-c_c)\sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_3 \\ &\quad - \sqrt{c_c(2-c_c)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3})^{-1}H_3 + o(\varepsilon_{1:s-1},H_1,H_3) \\ &= (1-c_1-c_\mu)^{-1}(1-c_c)^2d^p_{j+1}H_1 + d^p_{j+3}H_3 + o(\varepsilon_{1:s-1},H_1,H_3) \ , \end{aligned}$$

where $d^p_k \coloneqq (1-c_1-c_\mu)^{-1/2}\left[1-c_c-\Gamma((1-c_\sigma)^k p)^{-1}\right]\sqrt{c_c(2-c_c)\mu_{\text{eff}}}$ for $k = j+1, j+3$. Then,

$$\begin{aligned} q^h_T = q^h_{T_0+j+5} &= (r^h_{T_0+j+4})^{-1/2}(1-c_c)q^h_{T_0+j+4} + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}H_5 \\ &= (1-c_1-c_\mu)^{-3/2}(1-c_c)^3d^p_{j+1}H_1 \\ &\quad + (1-c_1-c_\mu)^{-1/2}(1-c_c)d^p_{j+3}H_3 + O(H_5) + o(\varepsilon_{1:s-1},H_1,H_3) \ . \end{aligned}$$

For $t = 0, \dots, T$, we denote $\boldsymbol{\Sigma}_t = R(\hat{\boldsymbol{\Sigma}}_t)^{-1}\hat{\boldsymbol{\Sigma}}_t$ and $\boldsymbol{\Sigma}^h_t = R(\hat{\boldsymbol{\Sigma}}^h_t)^{-1}\hat{\boldsymbol{\Sigma}}^h_t$. For $t = T_0, \dots, T-1$, given the choice of $v_{T_0+1:T_0+j+5} = 0$ in (B.31), we have then $\hat{\boldsymbol{\Sigma}}_{t+1} = (1-c_1-c_\mu)\boldsymbol{\Sigma}_t/\rho((1-c_1-c_\mu)\boldsymbol{\Sigma}_t)$ and by **R**1

$$\boldsymbol{\Sigma}_{t+1} = \frac{\hat{\boldsymbol{\Sigma}}_{t+1}}{R(\hat{\boldsymbol{\Sigma}}_{t+1})} = \frac{(1-c_1-c_\mu)\boldsymbol{\Sigma}_t}{R((1-c_1-c_\mu)\boldsymbol{\Sigma}_t)} = \boldsymbol{\Sigma}_t \ .$$

Thus, $\mathbf{\Sigma}_t = \mathbf{\Sigma}_T$ for $t = T_0, \dots, T$. Moreover, we have for $k = 0, \dots, j$:

$$\hat{\mathbf{\Sigma}}^h_{T_0+k+1} = \frac{(1-c_1-c_\mu)\mathbf{\Sigma}^h_{T_0+k} + c_1 q^h_{T_0+k+1}(q^h_{T_0+k+1})^\top + c_\mu \sum_{i=1}^\mu w_i^c h^i_{T_0+k+1}(h^i_{T_0+k+1})^\top}{\rho((1-c_1-c_\mu)\mathbf{\Sigma}^h_{T_0+k} + c_1 q^h_{T_0+k+1}(q^h_{T_0+k+1})^\top + c_\mu \sum_{i=1}^\mu w_i^c h^i_{T_0+k+1}(h^i_{T_0+k+1})^\top)}$$

Since ρ is $\mathcal{C}^1$ by $\boldsymbol{\rho}2$, hence locally Lipschitz, and $q^h_{T_0+k+1}(q^h_{T_0+k+1})^\top = 0 + O(\|h_{1:T_0+k}\|^2) = o(h_{1:T_0+k})$, we have then:

$$\hat{\mathbf{\Sigma}}^h_{T_0+k+1} = \frac{(1-c_1-c_\mu)\mathbf{\Sigma}^h_{T_0+k}}{\rho((1-c_1-c_\mu)\mathbf{\Sigma}^h_{T_0+k})} + o(h_{1:T_0+k}) = \frac{\mathbf{\Sigma}^h_{T_0+k}}{\rho(\mathbf{\Sigma}^h_{T_0+k})} + o(h_{1:T_0+k}) = \hat{\mathbf{\Sigma}}^h_{T_0+k} + o(h_{1:T_0+k}) \ ,$$

where we have used $\boldsymbol{\rho}1$ to simplify the above equation. Therefore, we obtain by induction and using (B.28) that

$$\hat{\mathbf{\Sigma}}^h_{T_0+k} = \hat{\mathbf{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}, H_1, H_3, H_5)$$

and thus, using $\mathbf{\Sigma}_{T_0+k} = \mathbf{\Sigma}_T$, and since R is locally Lipschitz by $\mathbf{R}2$:

$$\begin{aligned}\mathbf{\Sigma}^h_{T_0+k} = \frac{\hat{\mathbf{\Sigma}}^h_{T_0+k}}{R(\hat{\mathbf{\Sigma}}^h_{T_0+k})} &= \frac{\hat{\mathbf{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}, H_1, H_3, H_5)}{R(\hat{\mathbf{\Sigma}}_{T_0} + \sum_{k=1}^{s-1} \varepsilon_k \mathbf{S}_k + o(\varepsilon_{1:s-1}, H_1, H_3, H_5))} \\ &= \underbrace{\frac{\hat{\mathbf{\Sigma}}_{T_0}}{R(\hat{\mathbf{\Sigma}}_{T_0})}}_{=\mathbf{\Sigma}_{T_0}} + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) = \mathbf{\Sigma}_T + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) \ . \quad \text{(B.35)}\end{aligned}$$

It follows that:

$$\begin{aligned}p^h_{T_0+j+1} &= (1-c_\sigma)p^h_{T_0+j} + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\mathbf{\Sigma}^h_{T_0+j})^{-1/2} \times (0 + H_1) \\ &= (1-c_\sigma)^{j+1}p + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{\Sigma}_T^{-1/2}H_1 + O(\varepsilon_{1:s-1}) + o(H_1)\end{aligned}$$

and

$$\begin{aligned}p^h_{T_0+j+2} &= (1-c_\sigma)p^h_{T_0+j+1} \\ &\quad + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\mathbf{\Sigma}^h_{T_0+j+1})^{-1/2} \times (0 - (1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}H_1) \\ &= (1-c_\sigma)^{j+1}p + (1-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{\Sigma}_T^{-1/2}H_1 \\ &\quad - \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)^{-1}\mathbf{\Sigma}_T^{-1/2}H_1 \\ &\quad + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) \\ &= (1-c_\sigma)^{j+2}p + c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H5) \ .\end{aligned}$$

Likewise,

$$\begin{aligned}p^h_{T_0+j+3} &= (1-c_\sigma)p^h_{T_0+j+2} + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\mathbf{\Sigma}^h_{T_0+j+2})^{-1/2} \times (0 + H_3) \\ &= (1-c_\sigma)^{j+3}p + (1-c_\sigma)c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 \\ &\quad + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{\Sigma}_T^{-1/2}H_3 + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5)\end{aligned}$$

and

$$\begin{aligned}p^h_{T_0+j+4} &= (1-c_\sigma)p^h_{T_0+j+3} \\ &\quad + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\mathbf{\Sigma}^h_{T_0+j+3})^{-1/2} \times (0 - (1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3}p)^{-1}H_3) \\ &= (1-c_\sigma)^{j+4}p + (1-c_\sigma)^2 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 + (1-c_\sigma)\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}\mathbf{\Sigma}_T^{-1/2}H_3 \\ &\quad - \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3}p)^{-1}\mathbf{\Sigma}_T^{-1/2}H_3 \\ &\quad + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) \\ &= (1-c_\sigma)^{j+4}p + (1-c_\sigma)^2 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 + c^p_{j+3}\mathbf{\Sigma}_T^{-1/2}H_3 + O(\varepsilon_{1:s-1}) + o(H_1, H_3, H_5) \ ,\end{aligned}$$

where $c_k^p \coloneqq (1-c_\sigma-(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^k p))^{-1}\sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}$ for $k=j+1, j+3$. Finally,

$$\begin{aligned} p^h_{T_0+j+5} &= (1-c_\sigma)p^h_{T_0+j+4} + \sqrt{c_\sigma(2-c_\sigma)\mu_{\text{eff}}}(\mathbf{\Sigma}^h_{T_0+j+4})^{-1/2}\times(0+H_5) \\ &= (1-c_\sigma)^{j+5}p + (1-c_\sigma)^3 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 \\ &\quad + (1-c_\sigma)c^p_{j+3}\mathbf{\Sigma}_T^{-1/2}H_3 + O(\varepsilon_{1:s-1}) + o(H_1,H_3) + O(H_5) \ . \end{aligned}$$

By **R**2, we have

$$\begin{aligned} r^h_T &= R\left((1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}^h_{T-1})^{-1}\hat{\mathbf{\Sigma}}^h_{T-1} + o(H_1,H_3,H_5)\right) \\ &= (1-c_1-c_\mu)R(\hat{\mathbf{\Sigma}}^h_{T-1})^{-1}R(\hat{\mathbf{\Sigma}}^h_{T-1}) + o(H_1,H_3,H_5) \\ &= 1-c_1-c_\mu + o(H_1,H_3,H_5) = r_T + o(H_1,H_3,H_5) \ , \end{aligned}$$

where we have used **R**1 to simplify the first line into the second line in the above equation. All in all, when $H_1, H_3, H_5, \varepsilon_{1:s-1} \to 0$,

$$\begin{aligned} \theta^h_T = \theta_T &+ \begin{bmatrix} O(\varepsilon_{1:s-1}) \\ O(\varepsilon_{1:s-1}) \\ 0 \\ \sum_{t=1}^{s-1}\varepsilon_t \mathbf{S}_t \\ 0 \end{bmatrix} + o(\varepsilon_{1:s-1}, H_1, H_3, H_5) \\ &+ \begin{bmatrix} (1-c_1-c_\mu)^{-1/2}\Gamma(p_T)^{-1}c_m H_5 \\ (1-c_\sigma)\times\left[(1-c_\sigma)^2 c^p_{j+1}\mathbf{\Sigma}_T^{-1/2}H_1 + c^p_{j+3}\mathbf{\Sigma}_T^{-1/2}H_3\right] + O(H_5) \\ (1-c_c)(1-c_1-c_\mu)^{-1/2}\times\left[(1-c_c)^2(1-c_1-c_\mu)^{-1}d^p_{j+1}H_1 + d^p_{j+3}H_3\right] + O(H_5) \\ 0 \\ 0 \end{bmatrix} \ . \end{aligned}$$

We identify the Taylor expansion of $S^T_{\theta_0}(v_{1:T}+h_{1:T})$ in (5.20), with $L_z = (1-c_1-c_\mu)^{-1/2}\Gamma(p_T)^{-1}c_m$ and $\mathbf{L}^{\mathbf{\Sigma}}_k = \mathbf{S}_k$ for $k=1,\dots,s-1$ and

$$\begin{aligned} \mathsf{W}_L = \text{span}(\xi^1_{1:T_0},\dots,\xi^{s-1}_{1:T_0})\times\{0\}^j \times &\begin{pmatrix} 1 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p) \end{pmatrix}^\top \mathbb{R}^{d\mu} \\ &\times \begin{pmatrix} 1 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+3}p) \end{pmatrix}^\top \mathbb{R}^{d\mu} \times \mathbb{R}^{d\mu} \end{aligned}$$

and $L\colon \mathsf{W}_L \to \mathbb{R}^{s-1}\times(\mathbb{R}^d)^3$ maps a vector $h_{1:T}\in\mathsf{W}_L$ to a vector $(\varepsilon_{1:s-1}, H_1, H_3, H_5) \in \mathbb{R}^{s-1}\times(\mathbb{R}^d)^3$ such that

$$h_{1:s-1} = \sum_{k=1}^{s-1}\varepsilon_k \xi^k_{1:T_0}; h_{s+j:s+j+1} = \begin{pmatrix} H_1 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)H_1 \end{pmatrix}$$

$$h_{s+j+2:s+j+3} = \begin{pmatrix} H_3 \\ -(1-c_1-c_\mu)^{-1/2}\Gamma((1-c_\sigma)^{j+1}p)H_3 \end{pmatrix}; h_{s+j+4} = H_5 \ .$$

Since the scalars $\varepsilon_1,\dots,\varepsilon_{s-1}\in\mathbb{R}$ and the vectors $H_1, H_3, H_5 \in \mathbb{R}^d$ above can be chosen arbitrary and independently of each other, the linear application $L\colon \mathsf{W}_L \to \mathbb{R}^{s-1}\times(\mathbb{R}^d)^3$ is surjective. □

B.3 Proof of Proposition 5.7

Proof of Proposition 5.7(a) and (c). Suppose that either $c_c \neq 1$, $c_\sigma \neq 1$ and $1-c_c \neq (1-c_\sigma)\sqrt{1-c_1-c_\mu}$, or that $c_c \neq 1$, $c_\sigma = 1$. Assume moreover that $c_\mu > 0$. Apply then Lemmas 5.11 and 5.12 to get that there exist $T\in\mathbb{N}$ and $v_{1:T} \in \overline{\mathcal{O}^T_{\theta_0}}$ with

(a) in the case $c_\sigma \neq 1$, rge $\mathcal{D}S^T_{\theta_0}(v_{1:T}) \supset \mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times \mathrm{T}_{\mathbf{\Sigma}_T}\rho^{-1}(\{1\})\times\{0\}$;

(c) in the case $c_\sigma = 1$, rge $\mathcal{D}S^T_{\theta_0}(v_{1:T}) \supset \mathbb{R}^d \times \{0\} \times \mathbb{R}^d \times \mathrm{T}_{\mathbf{\Sigma}_T}\rho^{-1}(\{1\}) \times \{0\}$.

In both cases, consider arbitrary $h_z, h_p, h_q \in \mathbb{R}^d$, $\mathbf{H}_\mathbf{\Sigma} \in T_{\mathbf{\Sigma}_T}\rho^{-1}(\{1\})$, with $h_p = 0$ if $c_\sigma = 1$, so that there exists $h_{1:T} \in \mathsf{V}^T$ satisfying $\mathcal{D}S^T_{\theta_0}(v_{1:T})h_{1:T} = (h_z, h_p, h_q, \mathbf{H}_\mathbf{\Sigma}, 0)^\top$. By Taylor expansion, we have then

$$S^T_{\theta_0}(v_{1:T} + h_{1:T}) = \begin{bmatrix} z_T + h_z \\ p_T + h_p \\ q_T + h_q \\ \hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma} \\ r_T \end{bmatrix} + o(h_{1:T}) \ .$$

Since R is positive and positively homogeneous, it is not constant around $\hat{\mathbf{\Sigma}}_T$. Besides, R is differentiable at $\hat{\mathbf{\Sigma}}_T$ and thus there exists $w \in \mathbb{R}^d$ such that $\mathcal{D}R(\hat{\mathbf{\Sigma}}_T)(ww^\top) \neq 0$. Consider the nonconstant smooth function

$$G_w \colon s \in \mathbb{R} \mapsto F_\mathbf{\Sigma}(q_t, R(\hat{\mathbf{\Sigma}}_t)^{-1}\hat{\mathbf{\Sigma}}_t, r_t; s[w, \dots, w]) \ ,$$

see (5.14). Since R is locally Lipschitz on $\mathcal{S}^d_{++}$, then R is locally Lipschitz on the submanifold rge G_w, which is nontrivial since G_w is nonconstant. Then, by Rademacher's theorem [19, Corollary B.5], R is differentiable at $G_w(s)$ for almost every $s \in \mathbb{R}$. Moreover, we know that $\hat{\mathbf{\Sigma}}_T = G_w(0)$ and that $\mathcal{D}R(\hat{\mathbf{\Sigma}}_T)(ww^\top) \neq 0$. Thus, by upper semicontinuity of Clarke's Jacobian [19, Proposition B.9], there exists a sufficiently small $s > 0$ such that $\mathcal{D}R(G_w(s))(ww^\top) \neq 0$.

Then, we can find $\epsilon = sw \in \mathbb{R}^d$ a nonzero vector small enough so that, if $v_{T+1} = [\epsilon, \dots, \epsilon] \in \overline{\mathcal{O}^1_{\theta_T}}$, then R is differentiable at $\hat{\mathbf{\Sigma}}_{T+1}$, and $\mathcal{D}R(\hat{\mathbf{\Sigma}}_{T+1})(\epsilon\epsilon^\top) \neq 0$. Moreover, up to taking $s > 0$ smaller, we can assume that Γ is differentiable at $p_{T+1} \neq 0$ by $\mathbf{\Gamma}1$. Hence by composition and by Lemma 5.9, $S^{T+1}_{\theta_0}$ is differentiable at $v_{1:T+1}$. Indeed, by chain rule [13, Corollary 2.6.6], we have

$$\mathcal{D}S^{T+1}_{\theta_0}(v_{1:T+1})h_{1:T+1} = \mathcal{D}F(S^T_{\theta_0}(v_{1:T}), v_{T+1})\left(\mathcal{D}S^T_{\theta_0}(v_{1:T})(h_{1:T}), h_{T+1}\right)$$

Let $h_{T+1} = [h, \dots, h] \in \mathbb{R}^{d\mu}$ for some arbitrary $h \in \mathbb{R}^d$. Then,

$$S^{T+1}_{\theta_0}(v_{1:T+1} + h_{1:T+1}) = \begin{bmatrix} F_z(z_T + h_z, p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\ F_p(p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma}); [\epsilon + h, \dots, \epsilon + h]) \\ F_q(q_T + h_q, r_T; [\epsilon + h, \dots, \epsilon + h]) \\ F_\mathbf{\Sigma}(q_T + h_q, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\ F_r(q_T + h_q, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma}), r_T; [\epsilon + h, \dots, \epsilon + h]) \end{bmatrix} + o(h_{1:T+1})$$

see (5.11)-(5.15). Moreover, we have

$$\begin{aligned} &F_z(z_T + h_z, p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\ &\quad = \frac{z_T + h_z + c_m(h + \epsilon)}{r^{1/2}_{T+1}\Gamma(p_{T+1}) + O(h, h_p, h_q, \mathbf{H}_\mathbf{\Sigma})} + o(h_{1:T+1}) \\ &\quad = F_z(z_T, p_T, R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + r^{-1/2}_{T+1}\Gamma(p_{T+1})^{-1}h_z + O(h, h_p, h_q, \mathbf{H}_\mathbf{\Sigma}) + o(h_{1:T+1}) \ , \end{aligned}$$

$$\begin{aligned} &F_p(p_T + h_p, R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{-1}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma}); [\epsilon + h, \dots, \epsilon + h]) \\ &\quad = (1 - c_\sigma)(p_T + h_p) + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\mathrm{eff}}}R(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{1/2}(\hat{\mathbf{\Sigma}}_T + \mathbf{H}_\mathbf{\Sigma})^{-1/2}(\epsilon + h) + o(h_{1:T+1}) \\ &\quad = F_p(p_T, R(\hat{\mathbf{\Sigma}}_T)^{-1}(\hat{\mathbf{\Sigma}}_T); [\epsilon, \dots, \epsilon]) + (1 - c_\sigma)h_p + O(h, \mathbf{H}_\mathbf{\Sigma}) + o(h_{1:T+1}) \ , \end{aligned}$$

$$\begin{aligned} &F_q(q_T + h_q, r_T; [\epsilon + h, \dots, \epsilon + h]) \\ &\quad = r_T^{-1/2}(1 - c_c)(q_T + h_q) + \sqrt{c_c(2 - c_c)\mu_{\mathrm{eff}}}(\epsilon + h) + o(h_{1:T+1}) \\ &\quad = F_q(q_T, r_T; [\epsilon, \dots, \epsilon]) + (1 - c_c)r_T^{-1/2}h_q + \sqrt{c_c(2 - c_c)\mu_{\mathrm{eff}}}h + o(h_{1:T+1}) = q_{T+1} + h^+_q + o(h_{1:T+1}) \ , \end{aligned}$$

where $h_q^+ = (1-c_c)r_T^{-1/2}h_q + \sqrt{c_c(2-c_c)\mu_{\text{eff}}}h$,

$$\begin{aligned}
&F_r(q_T + h_q, R(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}})^{-1}(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\
&\quad = R\left((1-c_1-c_\mu)R(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}})^{-1}(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}}) + c_1[q_{T+1} + h_q^+][q_{T+1} + h_q^+]^\top + c_\mu[\epsilon + h][\epsilon + h]^\top\right) \\
&\qquad + o(h_{1:T+1}) \\
&\quad = R(\tilde{\boldsymbol{\Sigma}}_{T+1}) + \mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})[(1-c_1-c_\mu)R(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}})^{-1}\mathbf{H}_{\boldsymbol{\Sigma}}] \\
&\qquad + \mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})[c_1[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top]] + \mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})[c_\mu[\epsilon h^\top + h\epsilon^\top]] + o(h_{1:T+1}) \\
&\quad = F_r(q_T, R(\hat{\boldsymbol{\Sigma}}_T)^{-1}(\hat{\boldsymbol{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + (1-c_1-c_\mu)R(\hat{\boldsymbol{\Sigma}}_T)^{-1}\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})\mathbf{H}_{\boldsymbol{\Sigma}} \\
&\qquad + c_1\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top] + c_\mu\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})[\epsilon h^\top + h\epsilon^\top] + o(h_{1:T+1}) \\
&\quad = F_r(q_T, R(\hat{\boldsymbol{\Sigma}}_T)^{-1}(\hat{\boldsymbol{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + \mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ + o(h_{1:T+1}) \ ,
\end{aligned}$$

where

$$\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ = (1-c_1-c_\mu)R(\hat{\boldsymbol{\Sigma}}_T)^{-1}\mathbf{H}_{\boldsymbol{\Sigma}} + c_1[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top] + c_\mu[\epsilon h^\top + h\epsilon^\top]$$

and

$$\tilde{\boldsymbol{\Sigma}}_{T+1} = (1-c_1-c_\mu)R(\hat{\boldsymbol{\Sigma}}_T)^{-1}(\hat{\boldsymbol{\Sigma}}_T) + c_1 q_{T+1}q_{T+1}^\top + c_\mu\epsilon\epsilon^\top \ .$$

Let $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. Since $c_\mu > 0$ and $\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})(\epsilon\epsilon^\top) \neq 0$ (since as seen above $\mathcal{D}R(\hat{\boldsymbol{\Sigma}}_{T+1})(\epsilon\epsilon^\top) \neq 0$ and $\hat{\boldsymbol{\Sigma}}_{T+1}$ is proportional to $\tilde{\boldsymbol{\Sigma}}_{T+1}$, see Lemma 5.8), there exists $h = l\varepsilon$, where

$$l = (y - c_1\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})[q_{T+1}x^\top + xq_{T+1}^\top] - (1-c_1-c_\mu)R(\hat{\boldsymbol{\Sigma}}_T)\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})\mathbf{H}_{\boldsymbol{\Sigma}})/(2c_\mu\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})(\epsilon\epsilon^\top)) \ ,$$

and $h_q = (1-c_c)^{-1}r_T^{1/2}(x - \sqrt{c_c(2-c_c)\mu_{\text{eff}}}l\epsilon)$ such that $h_q^+ = x$ and

$$\mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ = (1-c_1-c_\mu)DR(\tilde{\boldsymbol{\Sigma}}_{T+1})R(\hat{\boldsymbol{\Sigma}}_T)^{-1}\mathbf{H}_{\boldsymbol{\Sigma}} + c_1 DR(\tilde{\boldsymbol{\Sigma}}_{T+1})[q_{T+1}x^\top + xq_{T+1}^\top] + 2lc_\mu[\epsilon\epsilon^\top] = y$$

Therefore, the linear map $(h_q, h) \mapsto (h_q^+, \mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+)$ valued in $\mathbb{R}^d \times \mathbb{R}$ is surjective. Besides,

$$\hat{\boldsymbol{\Sigma}}_{T+1}^h \coloneqq F_{\boldsymbol{\Sigma}}(q_T + h_q, R(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}})^{-1}(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) = \frac{\tilde{\boldsymbol{\Sigma}}_{T+1}^h}{\rho(\tilde{\boldsymbol{\Sigma}}_{T+1}^h)} \ ,$$

where

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_{T+1}^h &= \tilde{\boldsymbol{\Sigma}}_{T+1} + (1-c_1-c_\mu)R(\hat{\boldsymbol{\Sigma}}_T)^{-1}\mathbf{H}_{\boldsymbol{\Sigma}} + c_1[q_{T+1}(h_q^+)^\top + h_q^+ q_{T+1}^\top] \\
&\quad + c_\mu[\epsilon h^\top + h\epsilon^\top] + o(h_{1:T+1}) = \tilde{\boldsymbol{\Sigma}}_{T+1} + \tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ + o(h_{1:T+1}) \ .
\end{aligned}$$

Therefore, by using the Taylor expansion $\rho(\tilde{\boldsymbol{\Sigma}}_{T+1}^h)^{-1} = \rho(\tilde{\boldsymbol{\Sigma}}_{T+1})^{-1} - \mathcal{D}\rho(\tilde{\boldsymbol{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ + o(h_{1:T+1})$ since ρ is positive and continuously differentiable by $\boldsymbol{\rho}2$, we get

$$\begin{aligned}
&F_{\boldsymbol{\Sigma}}(q_T + h_q, R(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}})^{-1}(\hat{\boldsymbol{\Sigma}}_T + \mathbf{H}_{\boldsymbol{\Sigma}}), r_T; [\epsilon + h, \dots, \epsilon + h]) \\
&\quad = F_{\boldsymbol{\Sigma}}(q_T, R(\hat{\boldsymbol{\Sigma}}_T)^{-1}(\hat{\boldsymbol{\Sigma}}_T), r_T; [\epsilon, \dots, \epsilon]) + \rho(\tilde{\boldsymbol{\Sigma}}_{T+1})^{-1}\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ - (\mathcal{D}\rho(\tilde{\boldsymbol{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+)\tilde{\boldsymbol{\Sigma}}_{T+1} + o(h_{1:T+1}) \ .
\end{aligned}$$

All in all, by Taylor expansion,

$$\mathcal{D}S_{\theta_0}^{T+1}(v_{1:T+1})h_{1:T+1} = \begin{bmatrix} r_{T+1}^{-1/2}\Gamma(p_{T+1})^{-1}h_z + O(h, h_p, h_q, \mathbf{H}_{\boldsymbol{\Sigma}}) \\ (1-c_\sigma)h_p + O(h, \mathbf{H}_{\boldsymbol{\Sigma}}) \\ h_q^+ \\ \rho(\tilde{\boldsymbol{\Sigma}}_{T+1})^{-1}\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ - (\mathcal{D}\rho(\tilde{\boldsymbol{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+)\tilde{\boldsymbol{\Sigma}}_{T+1} \\ \mathcal{D}R(\tilde{\boldsymbol{\Sigma}}_{T+1})\tilde{\mathbf{H}}_{\boldsymbol{\Sigma}}^+ \end{bmatrix} \ ,$$

which proves that every element in $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\boldsymbol{\Sigma}}_{T+1}}\rho^{-1}(\{1\}) \times \mathbb{R}$ is reached by the linear map $\mathcal{D}S_{\theta_0}^{T+1}(v_{1:T+1})$ when $c_\sigma \neq 1$ so $\mathcal{D}S_{\theta_0}^{T+1}(v_{1:T+1})$ is surjective, hence of maximal rank, which proves the statement (a) (with $T+1$ instead of T). When $c_\sigma = 1$, the statement (c) follows as well as there exists $p = 0 \in \mathbb{R}^d$ such that for every $(z, q, \boldsymbol{\Sigma}, r) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathrm{T}_{\hat{\boldsymbol{\Sigma}}_{T+1}}\rho^{-1}(\{1\}) \times \mathbb{R}$, $(z, p, q, \boldsymbol{\Sigma}, r)$ belongs to the range of $\mathcal{D}S_{\theta_0}^{T+1}(v_{1:T+1})$. □

References

[1] Esther Tolulope Aboyeji, Oladayo S. Ajani, and Rammohan Mallipeddi. Covariance matrix adaptation evolution strategy based on ensemble of mutations for parking navigation and maneuver of autonomous vehicles. *Expert Systems with Applications*, 249:123565, September 2024.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery.

[3] Youhei Akimoto, Anne Auger, Tobias Glasmachers, and Daiki Morinaga. Global Linear Convergence of Evolution Strategies on More than Smooth Strongly Convex Functions. *SIAM Journal on Optimization*, 32(2):1402–1429, June 2022.

[4] D.V. Arnold and H.-G. Beyer. Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622, April 2004.

[5] Anne Auger. Convergence results for the $(1, \lambda)$-sa-es using the theory of ϕ-irreducible markov chains. *Theoretical Computer Science*, 334(1-3):35–69, 2005.

[6] Anne Auger. Analysis of Comparison-based Stochastic Continuous Black-Box Optimization Algorithms. Thèse d'habilitation à diriger des recherches, Université Paris-Sud, May 2016.

[7] Anne Auger and Nikolaus Hansen. Linear convergence on positively homogeneous functions of a comparison based step-size adaptive randomized search: the (1+ 1) es with generalized one-fifth success rule. *arXiv preprint arXiv:1310.8397*, 2013.

[8] Anne Auger and Nikolaus Hansen. Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains. *SIAM Journal on Optimization*, 26(3):1589–1624, January 2016.

[9] Jonathan Bieler, Rosamaria Cannavo, Kyle Gustafson, Cedric Gobet, David Gatfield, and Felix Naef. Robust synchronization of coupled circadian and cell cycle oscillators in single mammalian cells. *Molecular Systems Biology*, 10(7):739, July 2014.

[10] Alexis Bienvenüe and Olivier François. Global convergence for evolution strategies in spherical problems: Some simple proofs and difficulties. *Theoretical Computer Science*, 306(1):269–289, September 2003.

[11] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, May 2011.

[12] Alexandre Chotard and Anne Auger. Verifiable conditions for the irreducibility and aperiodicity of Markov chains by analyzing underlying deterministic models. *Bernoulli*, 25(1):112–147, February 2019.

[13] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, January 1990.

[14] Sebastian Colutto, Florian Fruhauf, Matthias Fuchs, and Otmar Scherzer. The CMA-ES on Riemannian Manifolds to Reconstruct Shapes in 3-D Voxel Images. *IEEE Transactions on Evolutionary Computation*, 14(2):227–245, April 2010.

[15] Lawrence Craig Evans and Ronald F Gariepy. *Measure Theory and Fine Properties of Functions, Revised Edition*. Chapman and Hall/CRC, New York, April 2015.

[16] Garuda Fujii, Youhei Akimoto, and Masayuki Takahashi. Exploring optimal topology of thermal cloaks by CMA-ES. *Applied Physics Letters*, 112(6):061108, February 2018.

[17] Marco A. Gallegos-Herrada, David Ledvinka, and Jeffrey S. Rosenthal. Equivalences of Geometric Ergodicity of Markov Chains. *Journal of Theoretical Probability*, May 2023.

[18] Armand Gissler. Evaluation of the impact of various modifications to CMA-ES that facilitate its theoretical analysis. In *GECCO 2023 - Genetic and Evolutionary Computation Conference*, July 2023.

[19] Armand Gissler, Alain Durmus, and Anne Auger. On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds, February 2024.

[20] Nikolaus Hansen, Dirk V Arnold, and Anne Auger. Evolution Strategies. 2015.

[21] Nikolaus Hansen and Anne Auger. CMA-ES: Evolution strategies and covariance matrix adaptation. In *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '11, pages 991–1010, New York, NY, USA, July 2011. Association for Computing Machinery.

[22] Nikolaus Hansen and Anne Auger. Principled Design of Continuous Stochastic Search: From Theory to Practice. In Yossi Borenstein and Alberto Moraglio, editors, *Theory and Principled Methods for the Design of Metaheuristics*, Natural Computing Series, pages 145–180. Springer, Berlin, Heidelberg, 2014.

[23] Nikolaus Hansen and Stefan Kern. Evaluating the CMA Evolution Strategy on Multimodal Test Functions. In *Parallel Problem Solving from Nature - PPSN VIII*, Lecture Notes in Computer Science, pages 282–291, Berlin, Heidelberg, 2004. Springer.

[24] Nikolaus Hansen, Sibylle D. Müller, and Petros Koumoutsakos. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, March 2003.

[25] Nikolaus Hansen and Andreas Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of IEEE International Conference on Evolutionary Computation*, pages 312–317, May 1996.

[26] Nikolaus Hansen and Andreas Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, June 2001.

[27] Nikolaus Hansen and Raymond Ros. Benchmarking a weighted negative covariance matrix update on the BBOB-2010 noiseless testbed. In *Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '10, pages 1673–1680, New York, NY, USA, July 2010. Association for Computing Machinery.

[28] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[29] Horn R. and Johnson C. *Matrix Analysis*. Cambridge University Press, 2013.

[30] G.A. Jastrebski and D.V. Arnold. Improving Evolution Strategies through Active Covariance Matrix Adaptation. In *2006 IEEE International Conference on Evolutionary Computation*, pages 2814–2821, July 2006.

[31] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2012.

[32] Atsuo Maki, Naoki Sakamoto, Youhei Akimoto, Hiroyuki Nishikawa, and Naoya Umeda. Application of optimal control theory based on the evolution strategy (CMA-ES) to automatic berthing. *Journal of Marine Science and Technology*, 25(1):221–233, March 2020.

[33] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[34] S. P. Meyn and P. E. Caines. Asymptotic Behavior of Stochastic Systems Possessing Markovian Realizations. *SIAM Journal on Control and Optimization*, 29(3):535–561, May 1991.

[35] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, December 2012.

[36] Daiki Morinaga, Kazuto Fukuchi, Jun Sakuma, and Youhei Akimoto. Convergence Rate of the (1+1)-ES on Locally Strongly Convex and Lipschitz Smooth Functions. *IEEE Transactions on Evolutionary Computation*, 28(2):501–515, April 2024.

[37] Andreas Ostermeier, Andreas Gawelczyk, and Nikolaus Hansen. Step-size adaptation based on non-local use of selection information. In Yuval Davidor, Hans-Paul Schwefel, and Reinhard Männer, editors, *Parallel Problem Solving from Nature — PPSN III*, pages 189–198, Berlin, Heidelberg, 1994. Springer.

[38] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(none):20–71, January 2004.

[39] Maria Rodriguez-Fernandez, Pedro Mendes, and Julio R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, 83(2):248–265, February 2006.

[40] Denis Serre. *Matrices: Theory and Applications*, volume 216 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2010.

[41] Takumi Tanabe, Kazuto Fukuchi, Jun Sakuma, and Youhei Akimoto. Level generation for angry birds with sequential VAE and latent variable evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '21, pages 1052–1060, New York, NY, USA, June 2021. Association for Computing Machinery.

[42] Cheikh Toure, Anne Auger, and Nikolaus Hansen. Global linear convergence of evolution strategies with recombination on scaling-invariant functions. *Journal of Global Optimization*, 86(1):163–203, May 2023.

[43] Cheikh Toure, Armand Gissler, Anne Auger, and Nikolaus Hansen. Scaling-invariant Functions versus Positively Homogeneous Functions. *Journal of Optimization Theory and Applications*, 191(1):363–383, October 2021.